JOINT INFERENCE OF HUMAN GENOMIC FUNCTION AND SELECTIVE PRESSURE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Brad Gulko

August 2017

JOINT INFERENCE OF HUMAN GENOMIC FUNCTION AND SELECTIVE PRESSURE

Brad Gulko, Ph. D.

Cornell University 2017

Selective pressure and molecular phenotype provide complimentary perspectives on functional properties of the human genome. In this dissertation, I develop computational methods for identifying collections of molecular phenotypes called *functional classes*, that are optimally informative about recent selective pressure in humans. Aggregating selective pressure across genomic positions within each functional class produces a score representing the probability that a position evincing a class-associated molecular phenotype is under selective pressure. A class's score is interpreted as a measure of potential for fitness-influencing genomic function. Functional classes and attendant selective pressure scores are developed over the course of two papers.

In the first paper, I investigate three ENCODE cell-types and develop a non-parametric representation of covariates from four genomic properties including DNase-seq, RNA-seq, chromatin state, and protein coding annotation. The resultant 624 classes and attendants scores, are shown to predict eQTL, transcription factor binding, and enhancers as well as or better than contemporary methods using high dimensional functional covariates, or selective constraint alone. The interpretation of the score as selective pressure is also shown to be consistent with previous measures of genome-wide selective pressure.

In the second paper, I expand the cell-type cohort to 115 Epigenomic Roadmap

cell-types and nine genomic properties including splicing, transcription factor binding, and small RNA-seq. Complexity constraints are developed to reduce the number of functional classes from more than 1.2 million possibilities to 61. The resultant functional classes, genomic segmentations, and positional scoring (*FitCons2* scores) are used to detect small features including disease associated variation from HGMD and ClinVar clinical databases. FitCons2 scores are shown to have power comparable or superior to contemporary methods designed specifically to detect such features. Functional classes and scores are also shown to identify cell-type specific regulatory behavior of promoters and enhancers, while highlighting regulatory relationships between differing cell-types and developmental stages. I demonstrate how cell-type sensitivity in FitCons2 scores can be used to address an unsolved biological problem in characterizing transcription factor binding in craniofacial enhancers that are believed to differentiate neural development between humans, chimpanzees, and Neanderthals.

BIOGRAPHICAL SKETCH

Brad Gulko wrote his first piece of software in the PILOT programming language from teletype machine at the age of 9. By16 he was developing software models for bond finance and supported himself during his lower division coursework at UCLA by developing classified nuclear burst simulations at California Research and Technology, a U.S. Department of Energy contractor. In 1992, Brad graduated from the University of California at Santa Cruz with baccalaureate majors in Mathematics and Physics and a focus on semi-supervised learning systems. While at UCSC, Brad came under the tutelage of professor David Haussler whose warmth and infectious enthusiasm inspired a focus on the nascent field of computational genomics. In 1995 Brad completed his MS in Computer Engineering specializing in the use of machine learning to model structural and evolutionary properties of nucleic acids. During summers, Brad's software development work for technology companies in nearby Silicon Valley provided access to large distributed networks of workstations. His nocturnal use of these machines as improvised clusters to complete his master's work, foreshadowed the distributed genome assembly process implemented in his UCSC lab to complete the draft reference first of the human genome in 2000.

In 2009 Brad began PhD studies in the Computer Science Department at Cornell University. After a brief but highly influential period studying decision theory and causal inference with computer science professor Joe Halpern, Brad joined Adam Siepel's computational genomic lab to work on the problem of integrating strongly diverse types of information to identify genomic regulatory mechanisms using evolutionary selective pressure in humans. He hopes to develop his PhD work into a regulatory lexicon as a step towards a more comprehensive theory of nonpathogenic biology in evolving regulatory networks.

For the many teachers in my life who showed me that consistency, honesty, and compassion could fruitfully coexist, that synergy was possible, and who ultimately gave me the opportunity to attempt this work.

For my family who knew me well, and supported my efforts anyway.

For my wife, Leanne, who married me to join this adventure knowing that I would soon return to graduate school and that we might not see it done until well after we were 50. Your bravery and determination are an inspiration.

ACKNOWLEDGMENTS

This work was made possible by the intentional help and gracious support of many, but also unwitting contributions of even more; researchers and students who toiled to produce and then make public vast oceans of genomic data without knowing exactly if, or how, it would eventually be used. To those researchers I know, and those I will never meet, thank you.

In particular, I would like to acknowledge the generous personal, financial, and academic support of my adviser Dr. Adam Siepel. Dr. Siepel's vision directly inspired my work, his advice kept me on track, his experience helped realistically constrain my research agenda and walk me through questions I was too inexperienced to ask. I am confident that without his assistance I would have never even made it to the starting point. From one Haussler Lab alum to another, thank you.

I would like to thank the members and alums of the Siepel lab who maintained collegiality, support, insight, and kindness during a difficult process. In particular: Dr. Charles Danko for unexpected mentorship, Dr. Ilan Gronau for frequent, productive, and occasionally heated discussion during the early part of this project (and of course, INSIGHT), Noah Dukler for energy, contemporary sensibilities, and the experience of a 7-day paper, Elizabeth Hutton for frequent and thoughtful feedback, Dr. Yifei Huang for encyclopedic knowledge, friendship, and ready analytical sparring, Melissa Hubisz for knowing when a crucial figure was just terrible, Ritika Ramani for spending endless days turning my data into something someone might want to see, and Talitha Forcier for a precious humanity in the long lab evenings.

I would also like to thank the other members of my PhD committee, Dr. Haiyuan Yu, Dr. Martin Wells, and Dr. Thorsten Joachims for supporting my work, and especially for reassurance and important conversations and direction. I'll never

look at an evolving binding domain, canonical correlation, or a discriminative learning system in quite the same way. I would also like to thank Computer Science Professor Dr. Joe Halpern, whose teachings on uncertainty, decision theory, and thoughtful logical conditioning, quite unexpectedly, inhabit every corner and shadow of this work.

My enduring appreciation all my to coauthors during my PhD program, especially those I have not yet mentioned: Dr. Samantha Leung, Dr. Leonardo Arbiza, Dr. Bulent Aksoy, Dr. Alon Keinan. Publication is difficult act of creation. Synergy among coauthors makes a seemingly impossible task seem suddenly plausible, then actually achievable. Thank you for your efforts, talents, and generosity in sharing your work with me.

On a personal note, I would also like to thank Dr. Elspeth Golden and Dr. Michael Fleming, my dear friends who illuminated the PhD path for me. Of course, I would also like to thank my wife Leanne Nebenzahl whose patience, support, and valuable feedback on what to include (and *not* to include) in a presentation nurtured and supported my enthusiasm for this work over many years. This would not have happened without her.

Work in Dr. Adam Siepel's laboratory was supported by a David and Lucile Packard Fellowship for Science and Engineering and a grant from the NIH/NIGMS (R01 GM102192).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

A method for calculating probabilities of fitness consequences for point mutations across the human genome

(Gulko B, Hubisz M, Gronau I, Siepel A. 2015. *Nature Genetics* 47:3 276-283)

1.1 INTRODUCTION

During the past decade, two major developments—the emergence of massively parallel, ultra-cheap DNA sequencing technologies and the use of these technologies as digital readouts for functional genomic assays—have led to a profusion of data describing various features of genomes, epigenomes and transcriptomes[1,2]. However, investigators still have only rudimentary tools for integrating these diverse sources of information to obtain useful insights about genomic function and evolution. The limitations of current methods are particularly evident in the vast noncoding regions of eukaryotic genomes, which, despite recent progress[3–6], remain poorly annotated and understood. These limitations hamper progress in many areas, including molecular genetics, disease association and personalized medicine[7].

Many computational methods for the functional analysis of sequence data are based on the simple but profound observation that functionally important nucleotides tend to remain unchanged over evolutionary time because mutations at these sites generally reduce fitness and are therefore eliminated by natural selection[7–15]. A major strength of these conservation- or constraint-based approaches is that they sidestep thorny questions about the relationship between the outcomes of biochemical experiments and fitness-influencing functional roles[16–19] by getting at fitness directly through observations of evolutionary change. In essence, the ‘experiment’ considered

by these methods is the one conducted directly on genomes by nature over millennia, and the outcomes of interest are the presence or absence of fixed mutations.

These conservation-based methods, however, depend critically on the assumption that genomic elements are present at orthologous locations and maintain similar functional roles over relatively long evolutionary time periods. Evolutionary turnover may cause inconsistencies between sequence orthology and functional homology that substantially limit this type of analysis. Consequently, investigators have developed two major alternative strategies for the identification and characterization of functional elements. The first strategy is to augment information about interspecies conservation with information about genetic polymorphism[20–28]. The shorter evolutionary time scales associated with intraspecies variation make this approach more robust to evolutionary turnover and less sensitive to errors in alignment and orthology detection. Polymorphic sites tend to be sparse along the genome, however, so this approach requires some type of pooling of information across genomic positions, which can be problematic in the absence of high-quality genomic annotations. The second strategy is to forgo the use of evolutionary information and to instead predict functional roles from genomic data alone, typically with machine learning methods for supervised classification[29,30] or clustering followed by labeling based on known examples[31–33]. This approach has the limitation that it depends strongly on previously characterized elements, which in noncoding regions are typically few and perhaps unrepresentative of the genome.

In this report, we introduce a method for genomic analysis that combines many of the strengths of these polymorphism-based and functional genomic approaches. Like functional genomic methods, our approach groups genomic regions according to functional genomic fingerprints across multiple assays. Instead of relying on known examples for classification, however, we characterize each group by a probability of

mutational fitness consequences—or fitCons score—inferred from patterns of genetic variation. These fitCons scores are estimated using a recently developed statistical method, called Inference of Natural Selection from Interspersed Genomically Coherent Elements (INSIGHT), that contrasts patterns of polymorphism and divergence for a collection of dispersed genomic sites with those for nearby neutrally evolving sites, accounting for negative and positive selection[34]. Thus, the method integrates both evolutionary and functional data in characterizing the potential functional importance of genomic regions. We demonstrate that these fitCons scores are useful for visualization, for prediction of *cis* regulatory elements and for measurement of the global influence of recent natural selection across the genome.

1.2 RESULTS

1.2.1 General features of the prediction problem

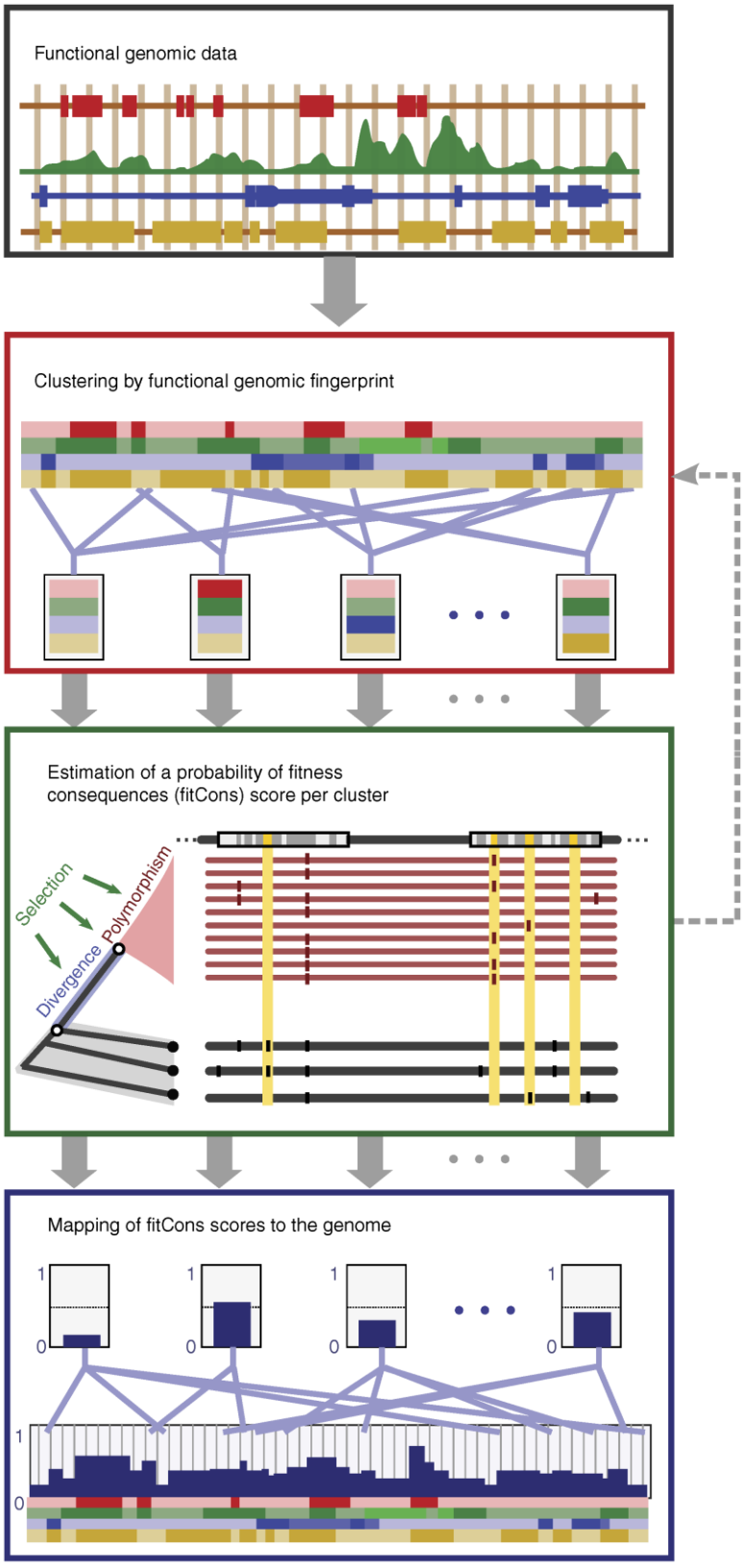
Information about genetic variation can be used to estimate probabilities of fitness consequences for moderately large groups of genomic positions but not for individual loci, owing to the sparsity of informative sites along the genome. This property of 'group-wise' but not 'individual' predictivity is common to many statistical problems, but it is complicated in our case by two additional features. First, an appropriate scheme for grouping or stratification is not clear a priori here because genomic correlates of fitness consequences are incompletely understood. Second, the outcomes of interest in our problem—the fitness consequences of point mutations—are not directly evident from the data. To highlight these challenges, consider the simpler problem of estimating the expected risk of an automobile accident. This problem must also be addressed at the level of groups (either explicitly, through stratification of drivers, or implicitly, through regression), but in this case the relevant features—such as the age, sex and number of traffic violations of the driver—are generally plain to the analyst. In addition, the outcomes of interest—the occurrences

and costs of accidents—are directly observed. In our problem, the genomic 'risk factors' for fitness-influencing mutations, particularly in unannotated noncoding regions of the genome, are much less clear. Furthermore, once a grouping is determined, it is still not possible to read off the associated fitness consequences of mutations; instead, they must be inferred from patterns of genetic variation using an evolutionary model.

1.2.2 Calculation of fitCons scores

We have addressed these challenges using the following strategy. Beginning with genome-wide functional genomic data sets obtained from each cell type (Fig. 1.1, first step), we first cluster genomic positions by their joint functional genomic fingerprints (Fig. 1.1, second step). We focus on three highly informative and largely orthogonal functional genomic data types—DNase I digestion and sequencing (DNase-seq) data, RNA sequencing (RNA-seq) data and chromatin immunoprecipitation and sequencing (ChIP-seq) data describing histone modifications—which describe DNA accessibility, transcription and chromatin states, respectively. We divide genomic positions into 3 levels of DNase-seq signal, 4 levels of RNA-seq signal and 26 distinct chromatin states on the basis of the ChromHMM method[31,33]. In addition, we distinguish between sites that fall outside or within annotated protein-coding sequences (CDSs). We then consider all possible combinations of these 4 types of assignments, obtaining $3 \times 4 \times 26 \times 2 = 624$ distinct functional genomic classes. We apply this clustering step separately to three karyotypically normal cell types: human umbilical vein epithelial cells (HUVECs), H1 human embryonic stem cells (H1 hESCs) and lymphoblastoid cells (GM12878), resulting in 443–447 usable classes of sites with median numbers of 165,000 to 224,000 sites per class (see Supplementary Table 1.1 and the Online Methods for details).

Figure 1.1: Procedure for calculating fitCons scores. Functional genomic data, such as DNase-seq, RNA-seq and histone modification data, are arranged along the genome sequence in tracks (first panel). Nucleotide positions in the genome are clustered by joint patterns across these functional genomic tracks (second panel). For example, one cluster might contain genomic positions with a high DNase-seq signal, a moderate RNA-seq signal and high signals for monomethylation of histone H3 at lysine 4 (H3K4me1) and acetylation of histone H3 at lysine 27 (H3K27ac), suggesting transcribed enhancers. Another might contain positions with a low DNase-seq signal, a high RNA-seq signal and a signal for trimethylation of histone H3 at lysine 36 (H3K36me3), suggesting actively transcribed gene bodies. Note that clusters will generally contain genomic positions dispersed along the genome sequence. Patterns of polymorphism and divergence are analyzed using INSIGHT[34] to obtain an estimate of the fraction of nucleotides under natural selection (ρ) in each cluster (third panel). This quantity is interpreted as the probability that each nucleotide position influences the fitness of the organism that carries it, or a fitness consequence (fitCons) score. The fitCons score for each cluster is assigned to all genomic positions that were included in the cluster (fourth panel). In this way, all nucleotide positions are assigned a score, but there can be no more distinct scores than there are clusters. Note that, in our initial work here, the clustering of genomic positions is accomplished by a simple exhaustive partitioning scheme that produces 624 distinct clusters. In future work, however, it may be desirable to iterate between clustering and calculating scores (dashed arrow).


Functional genomic data
Clustering by functional genomic fingerprint
Estimation of a probability of fitness
consequences (fitCons) score per cluster
Selection
Polymorphism
Divergence
Mapping of fitCons scores to the genome
1
0

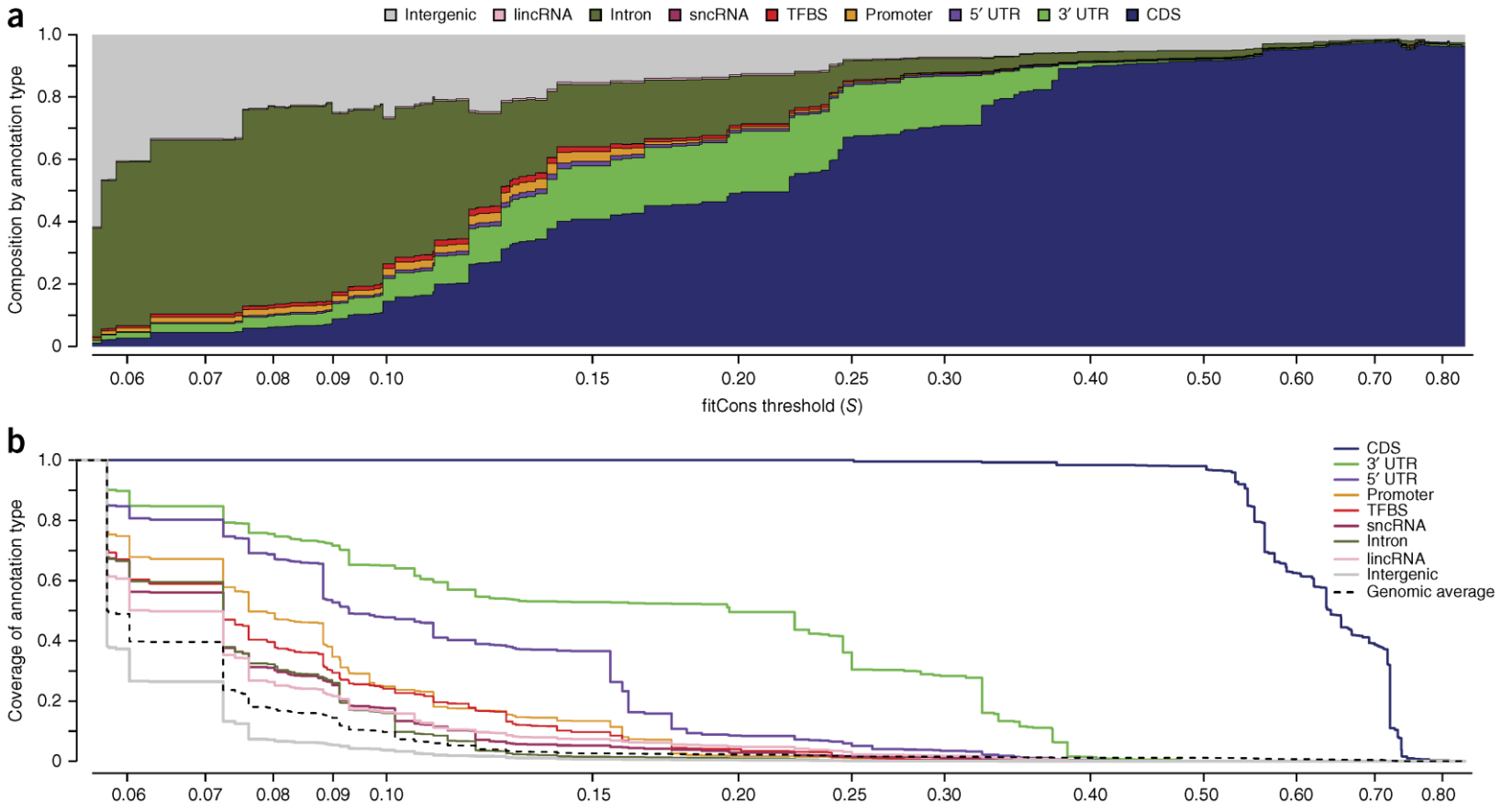
Next, we use INSIGHT to estimate the probabilities of mutational fitness consequences within each of these classes on the basis of patterns of polymorphism and divergence (Fig. 1.1, third step). This step yields an estimate of the fraction of sites under selection (ρ) for each of the analyzed classes, which serves as the fitCons score for that class. Finally, we assign to each nucleotide position in the genome the score estimated for the corresponding functional genomic class (Fig. 1.1, fourth step). Each genomic position is thus assigned a value between 0 and 1, representing the probability that the nucleotide at that position influences fitness, as estimated from patterns of variation at all genomic sites displaying the same functional genomic fingerprint. A vital property of these fitCons scores is that they integrate information from both evolutionary data and cell type–specific functional genomic data.

1.2.3 Genomic distribution of fitCons scores

To obtain a general overview of the genomic distribution of fitCons scores, we first considered the composition and coverage of nucleotide sites of various annotation types as a variable threshold S was applied to the fitCons score, focusing on HUVECs (see the Discussion for a summary of other cell types). When S is zero, all sites are considered and the composition of annotations reflects the overall genomic distribution (Fig. 1.2a). As S increases, however, sites in known functional classes become strongly enriched relative to intergenic and intronic sites. Regions such as 5′ and 3′ UTRs, promoters and introns are most enriched at intermediate scores, reflecting moderate levels of natural selection in these regions, whereas CDSs dominate at the highest scores. Coverage properties (Fig. 1.2b) are best for CDSs, 3′ UTRs and 5′ UTRs (in that order), but they are also considerably elevated above the intergenic background for promoters, transcription factor binding sites, long intergenic noncoding RNAs (lincRNAs) and small noncoding RNAs (sncRNAs). Notably, the enrichment for functionally annotated genomic regions at high scores occurs despite

no use of genomic annotations in the scoring scheme (except for CDS annotations). Instead, these elevated scores reflect differences in patterns of polymorphism and divergence that arise naturally from the fitness consequences of mutations in these regions and become evident after clustering on the basis of functional genomic data. The fitCons scores for each cell type are displayed across the genome as tracks in the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (Fig. 1.3 and Supplementary Fig. 1.7).

Figure 1.2: Composition and coverage of high scoring genomic regions according to fitCons. (a) Composition by annotation type in regions that exceed a fitCons score threshold of S, as S is varied across the range of possible scores. Each vertical cross-section of the plot can be thought of as a narrow 'stacked bar' representation of the composition by annotation type of all genomic positions at which the fitCons score is $>S$. At the left side of the plot, when S is small, the composition by annotation type is representative of the genome as a whole. As the threshold S increases, CDSs are increasingly enriched and intergenic sequences are increasingly depleted. Regions experiencing moderate levels of selection, such as UTRs, promoters, sncRNAs and introns, are most enriched at intermediate scores. Note the logarithmic scale for the x axis. TFBS, transcription factor binding site. (b) Coverage of the same annotation types by genomic regions having fitCons score $>S$, with an x axis matching that in (a). The dashed line indicates the genome-wide average. At each value of S, the relative height of a given curve in comparison to the dashed line indicates the enrichment (or depletion) of the corresponding annotation type in genomic regions having score $>S$. The legend at the right lists the annotation types in order of decreasing enrichment. When multiple annotations applied to a single nucleotide position, one was selected in the following order: CDS, transcription factor binding site, promoter, sncRNA, lincRNA, 5′ UTR, 3′ UTR, intron and intergenic. These figures summarize data at 2.9 billion genomic sites.


a
Intergenic
lincRNA
Intron
sncRNA
TFBS
Promoter
5′ UTR
3′ UTR
CDS
Composition by annotation type
1.0
0.8
0.6
0.4
0.2
0
0.06
0.07
0.08
0.09
0.10
0.15
0.20
0.25
0.30
0.40
0.50
0.60
0.70
0.80
fitCons threshold (S)
b
Coverage of annotation type
CDS
3′ UTR
5′ UTR
Promoter
TFBS
sncRNA
Intron
lincRNA
Intergenic
Genomic average

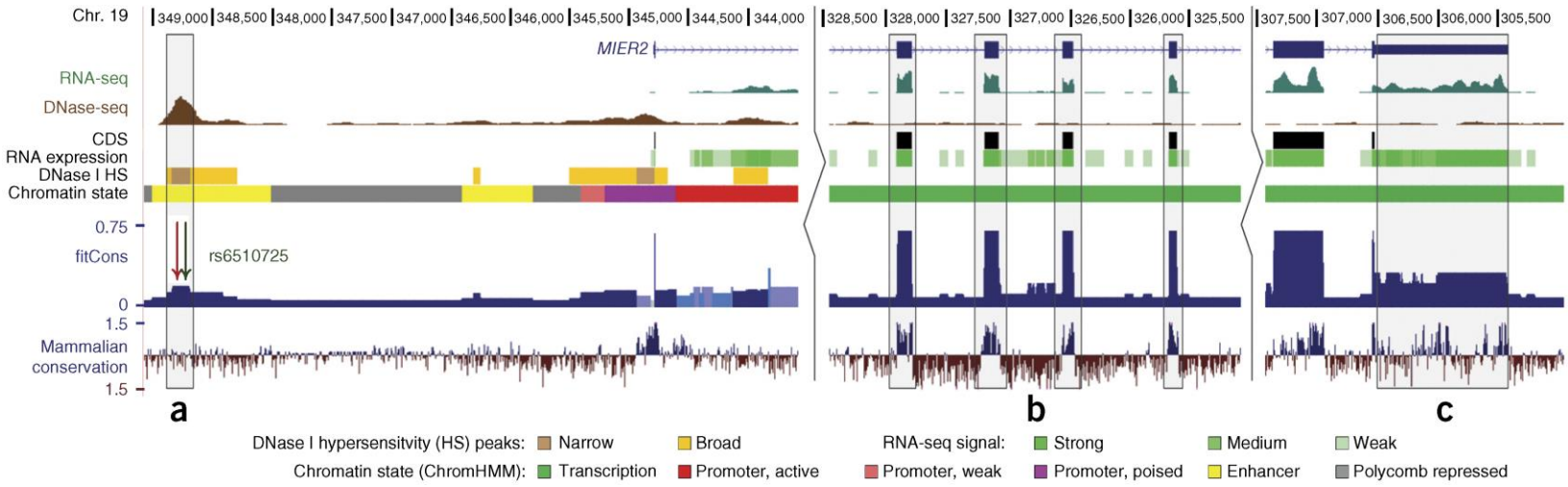


Figure 1.3: Genome browser display showing functional genomic fingerprints and fitCons scores. Shown, from top to bottom, are the exons of the MIER2 gene; the raw RNA-seq and DNase-seq signals; the 4 discretized tracks used to define the 624 functional genomic fingerprints, including annotation-based CDSs, RNA-seq signal, DNase-seq signal and chromatin modifications; the fitCons scores based on those fingerprints (dark blue, with lighter blues less statistically significant); and, for comparison, phyloP-based conservation scores for mammals. (a) An apparent enhancer, marked by a combination of enhancer-associated chromatin modifications and a strong DNase-seq signal, displays elevated fitCons scores but no elevation in conservation scores. Many regulatory elements display such a pattern, either because they have arisen recently in evolutionary time or because errors in orthology detection or alignment result in spuriously low conservation scores. Here a ChIP-seq–supported transcription factor binding site for AP-1 (red arrow) and a lung cancer–associated SNP (green arrow) are highlighted. (b) CDS exons show elevated scores according to both fitCons and phyloP. (c) The 3′ UTR, marked by transcription-associated chromatin modifications, a high RNA-seq signal and an absence of DNase I hypersensitivity or CDS annotations, displays moderately elevated fitCons scores and patches of evolutionary conservation. fitCons scores are fairly well correlated with phyloP conservation scores[15] across the genome, with some notable exceptions in noncoding regions (Supplementary Fig. 1). Browser tracks are publicly available on the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser (hg19 assembly).

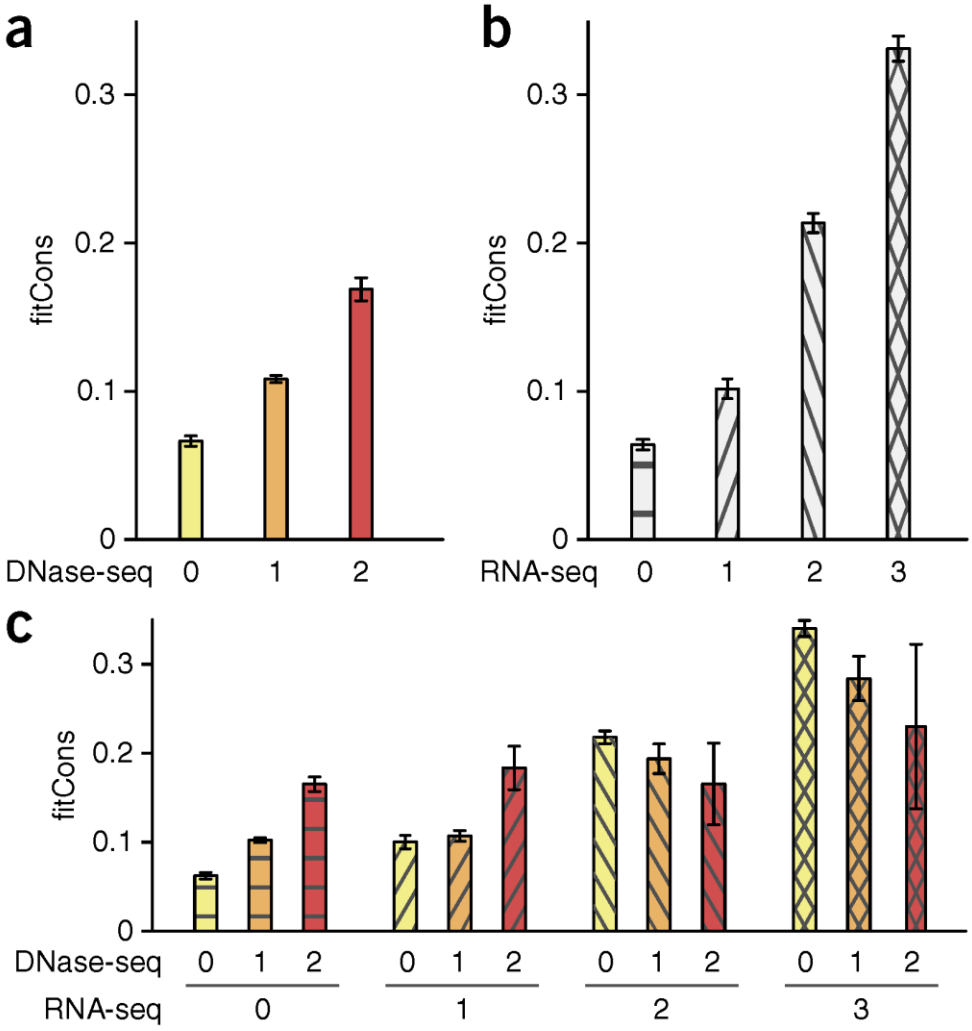


Figure 1.4: Average fitCons scores as a function of DNase-seq and RNA-seq intensity. Results represent averages across all non-CDS clusters having the marginal or joint property of interest. Error bars represent standard errors of the aggregated scores (Online Methods). (a) fitCons scores increase with DNase-seq intensity, probably owing to an increasing density of *cis* regulatory elements: 0, no DNase-seq signal; 1, broad peaks; 2, narrow peaks. (b) fitCons scores increase with RNA-seq intensity: 0, no RNA-seq reads; 1–3, weak to strong RNA-seq signal (Online Methods). (c) fitCons scores behave in a non-additive manner as joint combinations of DNase-seq and RNA-seq intensity are considered. In particular, at medium to high RNA-seq read depth (classes 2 and 3), fitCons scores decrease (rather than increase) with increasing DNase-seq signal. This unexpected pattern is explained by enrichment for DNase I hypersensitivity near the 5′ ends of genes. Conditional on a high RNA-seq signal, a high DNase-seq signal tends to be associated with the 5′ UTRs and upstream regions of genes, which are under fairly weak selection, whereas a low DNase-seq signal is associated with 3′ UTRs, which are under stronger selection. Each bar in (a) summarizes 104 clusters, each bar in (b) summarizes 78 clusters and each bar in c summarizes 26 clusters.

fitCons scores generally depend in expected ways on the marginal signals of functional genomic covariates, but they are also capable of capturing complex, non-additive relationships among covariates. For example, the scores outside of CDSs increase with marginal DNase-seq (Fig. 1.4a) and RNA-seq (Fig. 1.4b) signals, as expected; yet, a closer examination shows that the scores actually decrease with DNase-seq intensity in the presence of high RNA-seq intensity, owing to an implicit partitioning of 5′ and 3′ UTRs by DNase-seq data (Fig. 1.4c). This example demonstrates that our exhaustive partitioning scheme allows the method to capture unanticipated relationships between functional genomic covariates and natural selection.

1.2.4 Predictive power for *cis* regulatory loci

We evaluated the predictive power of fitCons scores for known cell type–specific regulatory elements in comparison with three widely used phylogenetic conservation scoring methods, the phastCons[12], phyloP[15] and Genomic Evolutionary Rate Profiling (GERP)[13] programs. In addition, we considered a new program, called Combined Annotation-Dependent Depletion (CADD)[35], that estimates the relative levels of pathogenicity of potential human variants using a support vector machine (SVM), many different genomic annotations and simulations of nucleotide divergence rates. Where appropriate, we also considered RegulomeDB, a scoring system for the regulatory potential of variant sites based on combined experimental and computational data[36], and EnhancerFinder, a kernel-based predictor for developmental enhancers based on multiple data types[37]. We evaluated the performance of these methods in predicting three types of functional elements that have putative roles in transcriptional regulation on the basis of different data sets: (i) binding sites for various transcription factors supported by ChIP-seq data from the Encyclopedia of DNA Elements (ENCODE) Project[3,28]; (ii) high-resolution expression quantitative

trait loci (eQTLs) identified in a recent large-scale study[6]; and (iii) enhancers identified on the basis of characteristic chromatin marks[38] (see the Online Methods for details).

To place the different predictors on equal footing, we plotted the base-wise coverage of each type of regulatory element as a function of the total coverage of the noncoding genome, varying score thresholds to include 0–20% of noncoding sites (Fig. 1.5). This strategy allowed us to measure the extent to which the elements of interest displayed signals that rose above the background of the noncoding genome, in a uniform manner across scoring methods. By this test, the fitCons scores showed dramatically better sensitivity for noncoding elements than almost all of the other methods considered. For example, at a total noncoding coverage of 2.5%, fitCons scores achieved nearly 70% coverage of transcription factor binding sites, whereas the other methods all had less than 20% coverage. Similarly, the coverage of enhancers was about 40% at 2.5% noncoding coverage, whereas most other scoring methods showed almost no signal above background. Only EnhancerFinder, which is specifically designed for this task, showed comparable prediction performance on enhancers. We also performed a more traditional evaluation of the tradeoff between sensitivity and specificity using receiver operating characteristic (ROC) curves and found that fitCons scores were considerably better predictors of regulatory function than all other methods considered (Online Methods and Supplementary Fig. 1.8).

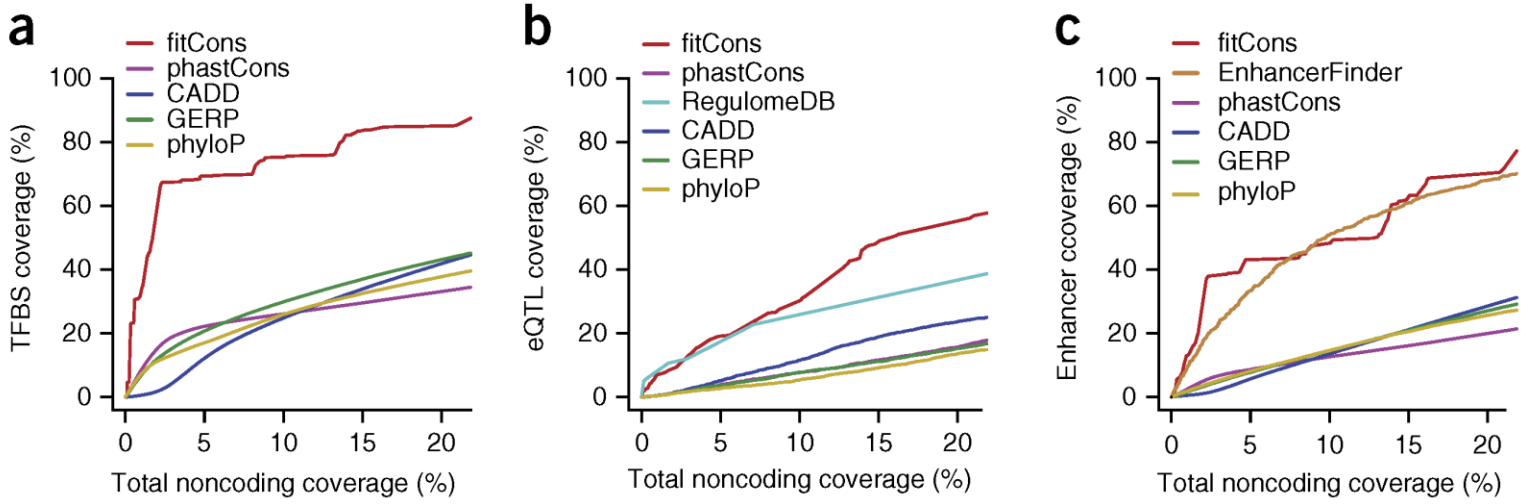


Figure 1.5: Coverage of active *cis* regulatory elements as a function of total coverage of the noncoding genome. Coverage of each type of element is shown as the score threshold is adjusted to alter the total coverage of noncoding sequences in the genome, excluding sites annotated as CDSs or UTRs. fitCons is compared with scores from the CADD[35], GERP[13], phastCons[12] and phyloP[15] programs (Online Methods). (a) Coverage of 55,844 transcription factor binding sites detected by ChIP-seq in HUVECs[28]. (b) Coverage of high-resolution eQTLs identified in a recent large-scale study[6], restricted to 3,662 eQTLs associated with genes transcribed in HUVECs. Coverage of eQTLs is also shown for classification of single-nucleotide variants by RegulomeDB[36]. The divergence-based scores (phastCons, phyloP, GERP and CADD) all perform poorly on the eQTL data set, probably because the ascertainment for segregating sites creates a bias against evolutionary conservation. Note also that the apparent performance of RegulomeDB, particularly at low total noncoding coverage, is somewhat influenced by consideration of eQTL data in its scoring scheme. (c) Coverage of 462 enhancers identified by characteristic chromatin marks[38] assayed in HUVECs. Coverage of these enhancers by EnhancerFinder[37] predictions is also shown. In all three plots, the x axis represents coverage at 2.8 billion noncoding positions.

The tests above were based on regulatory elements that are putatively active in the cell type for which the scores were produced, to highlight the benefits of using cell type–specific functional data. To evaluate how well these advantages extended across cell types, we created an integrated fitCons score by combining information from three cell types (Online Methods) and evaluated the performance of this score in predicting regulatory elements pooled from multiple cell types. We found that, in this less favorable setting, the fitCons scores still had better predictive performance for *cis* regulatory elements than any of the other scoring methods (Supplementary Fig. 1.9).

To address possible deficiencies of these tests, we carried out two additional sets of validation experiments. First, we performed a second round of experiments on ChIP-seq–supported transcription factor binding sites that considered only the subset of nucleotide positions at which base preferences were especially strong, which should be enriched for bases having fitness consequences. The ROC curves based on this more stringent test were very similar to the original curves (Supplementary Fig. 1.10), demonstrating that the apparent performance of the fitCons scores was not artificially inflated by the coarse-grained nature of our scores and transcription factor binding sites. Second, we examined an alternative set of predicted enhancers for GM12878 cells identified on the basis of characteristic patterns of divergent transcription initiation[39]. Unlike the chromatin-based enhancer predictions described above, these predictions were based on data completely independent from those underlying the fitCons scores. Nevertheless, the fitCons scores still displayed excellent predictive power for this set, better than all other methods besides EnhancerFinder (Supplementary Fig. 1.11).

1.2.5 Proportion of the human genome under selection

The proportion of nucleotides in the human genome that directly influence fitness—sometimes called the ‘share under selection’ (SUS)—has primarily been

estimated using methods that consider divergence patterns among mammals, for which turnover of functional elements might be an important confounding factor[40–44]. In addition to being useful as predictors of function, the fitCons scores could be useful in obtaining estimates of the SUS that are less sensitive to turnover because they measure natural selection over much shorter time scales.

An initial estimate of the SUS can be obtained by simply averaging the fitCons scores across all nucleotide positions in the genome. Because each score represents a probability that an individual nucleotide influences fitness, the average of these scores represents an expected fraction of nucleotides in the genome having fitness-influencing functions, or an expected SUS. This approach yielded an estimate of 7.5% (±0.1%) for HUVECs or 7.5–7.8% across the cell types. These estimates are largely consistent with but on the high end of those based on cross-species divergence, which generally have fallen between 3 and 8% (refs. 12,40,44–46). Among the sites under selection, we estimate that 9.0% are in CDSs, 2.2% are in 3′ UTRs, 35.2% are in introns, 51.7% are in intergenic regions and <1% are in each of several other noncoding annotation classes (Supplementary Table 1.2). Our estimates of the SUS are somewhat lower than previous estimates that have explicitly allowed for evolutionary turnover, most of which have been two to three times higher than the pan-mammalian estimates of ~5% (refs. 26,44,46–48). However, they are similar to a recent estimate of 7.1–9.2% based on improved alignments and a new model for turnover[49].

Violations of modeling assumptions will tend to bias fitCons scores upward, particularly for functional classes for which the true fraction is close to zero (Supplementary Note). To address this problem, we performed a parallel calculation for ‘neutral’ sites that intersected the large class of genomic positions having a ‘null’ functional genomic fingerprint (no DNase-seq, RNA-seq or histone modification

signal). This calculation resulted in an estimate of 3.3%, which can be considered an upper bound on the contribution of error because these putatively neutral sites undoubtedly include some sites under selection. By subtracting this 3.3% from our naive estimate of 7.5%, we obtained an estimated lower bound for the SUS of 4.2%, with somewhat higher fractions of selected sites in CDSs and 3′ UTRs (Supplementary Table 1.2). (These estimates are for HUVECs, but the results for the other cell types were very similar.) Overall, our analysis of the SUS suggests that between 4.2 and 7.5% of nucleotides in the genome have direct fitness-influencing functions and that the ratio of noncoding to coding functional sites is between 5.4 and 10.1.

1.2.6 Implications for evolutionary turnover of functional elements

To better understand the differences between fitCons scores and conventional divergence-based scores, we devised an alternative scoring system (denoted fitConsD) based on the same site clusters but an estimator of the fraction of nucleotides under selection that instead considers nucleotide divergence patterns across primates (Online Methods). Thus, the fitCons and fitConsD scores both represent probabilities of fitness consequences per nucleotide but over two different evolutionary time scales. Overall, these two measures were remarkably well correlated, with R^2 = 0.88 (Fig. 1.6a). Furthermore, a measure based on the difference between fitConsD and fitCons scores suggested relatively low amounts of turnover across annotation classes, accounting for no more than about 10% of all functional sites (Fig. 1.6b). These observations suggest that the main signal for selection has been maintained over long evolutionary time periods and that turnover has been modest during primate evolution but that there are some classes of sites that show stronger recent than ancient natural selection.

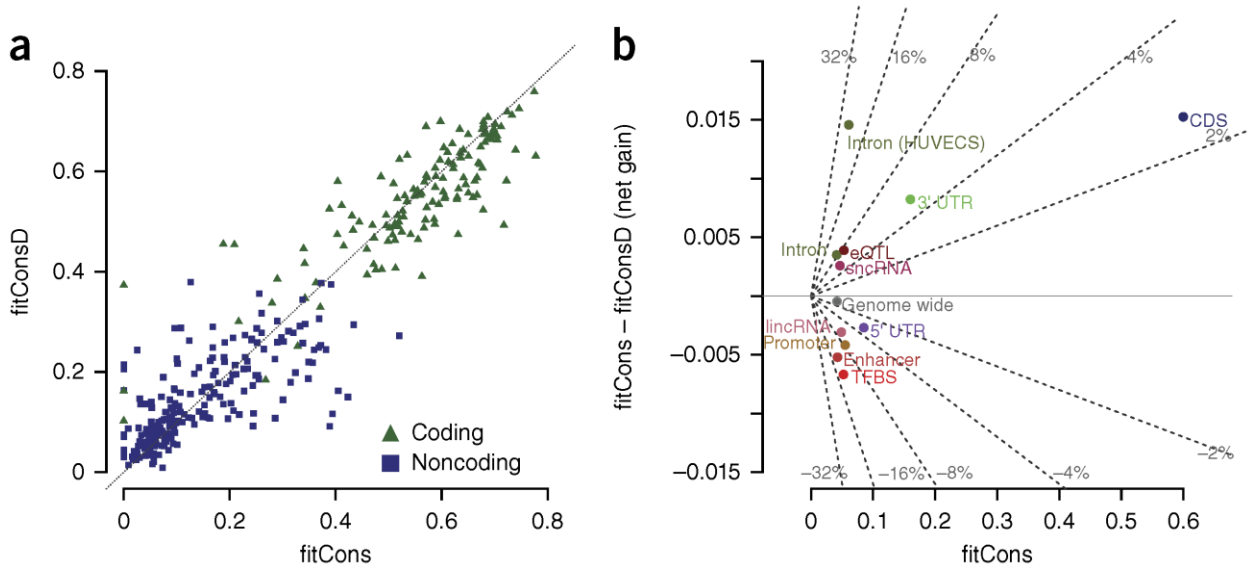


Figure 1.6: Comparison between fitCons and fitConsD scores. fitConsD is an alternative estimate of fitness consequences, analogous to fitCons but based on an estimator of the fraction of sites under natural selection that considers divergence patterns across four primate genomes (Online Methods). (a) fitCons and fitConsD scores are shown for the clusters defined using functional genomic data from HUVECs. Scores are shown for the 348 clusters of size 10 kb or larger, distinguishing between coding clusters (green triangles) and noncoding clusters (blue squares). Both sets of scores are corrected by subtracting the possible contribution from model misspecification (Online Methods). Correlation between the two sets of scores is high overall (R^2 = 0.88) and is somewhat higher for coding (R^2 = 0.69) than for noncoding (R^2 = 0.51) clusters. (b) The net gain in the fraction of sites under selection on population genetic time scales relative to primate divergence time scales, computed by subtracting average fitConsD scores from average fitCons score for different classes of functional elements (negative values imply net loss). Net gain is plotted against average fitCons score, and lines of constant slope radiating from the origin represent constant values of a 'net gain rate' per functional site, computed as NGR = (fitCons − fitConsD)/fitCons. The NGR is small (≤10%) for almost all annotation classes considered, with the main exception being the introns of active genes (NGR > 20%; see "Intron (HUVECs)"), which are enriched in clusters that exhibit an absence of DNase-seq or RNA-seq signal and chromatin modifications, suggesting transcriptional elongation.

1.3 DISCUSSION

The essential idea of our approach is to use functional genomic data to group sites into classes that are relatively homogeneous in terms of their functional roles, then to characterize the bulk influence of natural selection on these classes on the basis of their patterns of polymorphism and divergence. For our estimation of natural selection, we make use of a recently developed probabilistic model of evolution and efficient algorithms for genome-wide inference (INSIGHT). We interpret INSIGHT-based estimates of fractions of nucleotides under selection as probabilities that each nucleotide influences fitness, or fitness consequence (fitCons) scores. Even with a simple clustering scheme, these fitCons scores appear to be highly informative about genomic function.

According to our experiments, fitCons scores have excellent predictive performance for putative *cis* regulatory elements, outperforming several divergence-based methods (phastCons, phyloP, GERP and CADD) and one annotation-based method (RegulomeDB) by clear margins. They also performed slightly better in enhancer prediction than EnhancerFinder, a program specifically designed for this purpose, although it should be noted that EnhancerFinder was trained on other cell types. Notably, prediction performance does not appear to be sensitive to the choice of neutral sites used by INSIGHT (Supplementary Figs. 1.12 and 1.13). In part, the observed improvement in performance reflects the use of cell type–specific data (Fig. 1.5 and Supplementary Fig. 1.8), but fitCons scores also show a clear performance advantage when considering all annotated elements rather than just active ones (Supplementary Fig. 1.9). Thus, the approach of grouping genomic sites by functional genomic signatures and then measuring group-wise fitness consequences on the basis of patterns of genetic variation appears to offer real benefits for the prediction of

regulatory function, as compared with methods that consider either genetic divergence or functional genomic data alone.

Interestingly, the recently published CADD method performed no better on our tests than conventional conservation scores, despite reports by the authors of substantial advantages over phyloP, phastCons, GERP and other methods[35]. This inconsistency appears to reflect several important differences between our validation experiments and those they reported. First, our tests focused specifically on putative *cis* regulatory elements, whereas many of their tests considered a mixture of coding and noncoding elements. In particular, the ClinVar database, which figured prominently in their experiments, includes very few noncoding variants (~5% of pathogenic variants). Second, when Kircher *et al*. did consider noncoding regions, they generally did not distinguish between *cis* regulatory elements and sequences that more directly influence the structure and content of protein-coding transcripts, such as splice sites. CADD has a natural advantage with these variants owing to its use of gene annotations, whereas the annotation-free fitCons scores may perform better in completely unannotated regions of the genome. Finally, the tests by Kircher *et al*. that explicitly considered putative *cis* regulatory elements were limited to a few loci and examined only correlations with saturation mutagenesis experiments, irrespective of a prediction threshold. We view our ROC-type comparisons based on multiple independent genome-wide sets of elements as a more direct and comprehensive demonstration of predictive power for *cis* regulatory elements. In any case, the comparison of these two closely related yet distinct approaches helps to identify strengths and weaknesses of each and may lead to new ideas for improved methodologies.

A side benefit of our model-based approach is that the base-wise probabilities of fitness consequences lead in a straightforward manner to an estimate of the SUS in

the human genome. This estimate of the SUS reflects time scales since the divergence of humans and chimpanzees, about 4–6 million years ago, unlike conventional estimates based on tens or hundreds of millions of years of mammalian evolution. Nevertheless, our estimate of the SUS, at 4.2–7.5%, ends up being remarkably similar to those based on longer time scales, which have generally fallen between 3 and 8% (refs. [12,40–43,45,51]). It also overlaps with a recent estimate of 7.1–9.2% based on patterns of insertion and deletion and an explicit model of evolutionary turnover[49]. We take the general concordance of these estimates, both with one another and with our fitCons- and fitConsD-based estimates, as a strong indication that the SUS has remained quite low (probably <10%) over various time scales in mammalian evolution. This finding stands in contrast to estimates that ~80% of nucleotides may be functional, based on measures of 'biochemical activity' (ref. 3). However, it is important to bear in mind that these evolutionary and biochemical estimates reflect somewhat different definitions of function, and this may explain some of the difference between them[16,18,19]. For example, the fitCons- and conservation-based estimates (excluding those based on indels) generally represent the fractions of positions at which point mutations will have fitness consequences, but they do not account for sequences (such as spacer elements) that would have fitness consequences if deleted but not mutated (see the Supplementary Note for discussion).

Apart from the absolute fraction of functional DNA in the human genome is the question of how much the functional content of the genome has changed over time through gains and losses of functional elements. Several studies have estimated that such turnover could allow the current SUS in the human genome to be ~2–3 times larger than estimated from comparisons across mammals[26,46–48,48]. Indeed, these findings have been proposed to explain, in part, the discordance between evolution-based and biochemical estimates of the functional fraction of the genome[26,52,53].

However, most of these analyses have accounted for turnover using relatively crude methods, for example, by relying on an apparently near-linear relationship between pairwise divergence and the estimated SUS[46,47] or by estimating functional content from mean SNP densities or derived allele frequencies in genomic regions not conserved across mammals[26] (but see ref. 49 for an improved model). Our analysis is more direct, by comparing analogous divergence-based and polymorphism-based estimates of the SUS calculated from exactly the same clusters of nucleotide positions. In addition, our analysis focuses on primate evolution, rather than attempting to account for turnover across mammals, where factors such as alignment error, orthology detection and genomic rearrangement can be problematic. The similarity between our estimates based on polymorphism (fitCons) and divergence (fitConsD) strongly suggests that evolutionary turnover has been modest during primate evolution, as massive turnover would be expected to lead to a substantial downward bias in the divergence-based estimates. Our power experiments indicate that this observation is not an artifact of reduced sensitivity in the fitCons scores. Nevertheless, we cannot rule out the possibility that compensating gains and losses on very recent time scales maintain a similar SUS while substantially altering the genomic composition of functional sequences.

We have focused on HUVECs in this report, but we also generated fitCons scores for two other cell types (H1 hESCs and GM12878 cells). A comparison across cell types (Supplementary Note) indicated that the genomic positions assigned to each functional class differed substantially across cell types, but equivalently defined clusters had concordant fitCons scores in the different cell types (Supplementary Fig. 1.14). When cell type–specific scores were examined, elements active in that cell type displayed significantly higher scores than inactive elements. Moreover, particular elements had higher scores in cell types for which they were active than in cell types

for which they were inactive (Supplementary Fig. 1.15). Notably, we found that a set of integrated scores based on a simple, heuristic procedure (Online Methods) performed nearly as well as the cell type–specific scores in the target cell types but much better on elements from mismatched or pooled cell types (Supplementary Fig. 1.16). With more flexible and scalable clustering techniques, it may be possible to improve these methods by considering all cell types simultaneously, clustering sites by functional genomic fingerprints corresponding to multiple cell types and then producing a single set of scores reflecting these joint patterns. Such improvements, together with increases in the resolution and quality of the available functional genomic data, should result in improved power for the prediction of individual functional elements and refined estimates of the SUS.

1.4 JOURNAL DETAILS

This section includes additional Journal-specific material that appeared in the printed version of the publication.

1.4.1 URLs

Cold Spring Harbor Laboratory mirror of UCSC Genome Browser, http://genome-mirror.cshl.edu/; UCSC Genome Browser, http://genome.ucsc.edu/;

INSIGHT, http://compgen.cshl.edu/INSIGHT/;

GENCODE v15, ftp://ftp.sanger.ac.uk/pub/gencode/release_15/;

GERP, http://mendel.stanford.edu/SidowLab/downloads/gerp/;

CADD, http://cadd.gs.washington.edu/download;

RegulomeDB, http://regulome.stanford.edu/downloads/;

Gerstein laboratory ENCODE nets, http://encodenets.gersteinlab.org/;

European Bioinformatics Institute's E-GEUV-1 data set, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/.

1.4.2 Acknowledgments

We thank L. Arbiza for helpful discussions and assistance with early analyses and G. Cooper for constructive criticism of our validation experiments and comparisons with CADD. This research was supported by US National Institutes of Health grant GM102192, a David and Lucile Packard Fellowship for Science and Engineering (to A.S.) and a postdoctoral fellowship from the Cornell Center for Comparative and Population Genomics (to I.G.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

1.4.3 Author Contributions

I.G. and A.S. conceived the study framework. B.G. and I.G. performed the experiments. All authors analyzed the data. B.G., M.J.H. and I.G. developed analysis tools. B.G., I.G. and A.S. wrote the manuscript. I.G. and A.S. supervised the research.

1.5 METHODS SUMMARY

This Methods section is provided in the online full-text version of the publication, but not in the print version.

1.5.1 Functional genomic data

RNA-seq and DNase-seq data for HUVECs, H1 hESCs and GM12878 cells were downloaded from the UCSC Genome Browser. Chromatin states for the same three cell types were downloaded from the European Bioinformatics Institute's FTP site (see Supplementary Table 1.3). For DNase-seq, we considered two replicate experiments from University of Washington (UW) data for each cell type. However, only one UW replicate was available for H1-hESCs, so additional DNase-seq data for this cell line was obtained from Duke University. For each replicate DNase-seq experiment, we downloaded broad and narrow peak calls. For RNA-seq, we selected a single replicate from the Caltech poly(A)+ 75-bp paired-end read data, after

examining several alternative data sets. For chromatin states, we used the 25-state ChromHMM segmentation generated in December 2012 (ref. 33).

1.5.2 Clustering approach

We produced a separate partitioning for each cell type on the basis of the functional genomic data. The broad and narrow DNase-seq peaks were used to partition sites in the genome into three mutually exclusive classes: sites that fell in a narrow peak in both replicate experiments (class 2); sites that fell in a broad peak in at least one replicate and did not fall in a narrow peak in both replicates (class 1); and sites that fell outside of all called peaks (class 0). This three-level scheme allowed for both high sensitivity (class 1) and high specificity (class 2). For H1-hESCs, only one set of broad peak calls was available to define class 1. For the RNA-seq data, we partitioned sites in the genome into four mutually exclusive classes (0–3) on the basis of the number of reads aligned at each position. Read depth thresholds were set separately for each cell type through a process that aims to minimize the conditional entropy of concentrations of predicted sites under selection (Supplementary Note). Chromatin states were defined directly from the 25 states in ChromHMM, with a 26th state containing sites not assigned to any chromatin class. The Cartesian product of these partitions, together with the partition into coding and noncoding sequences, resulted in $3 \times 4 \times 26 \times 2 = 624$ distinct functional classes.

1.5.3 Running INSIGHT

The INSIGHT method infers the fraction of nucleotide sites under selection (ρ) for a given collection of sites by comparing patterns of within-species polymorphism and between-species divergence within these sites and within putatively neutrally evolving sites nearby. A detailed description of the method, sequence data and data quality filters is given in ref. 34. For each non-empty fitCons site cluster, INSIGHT was used to estimate ρ, which was used as the fitCons score of all sites in that cluster.

To reduce sensitivity to estimates with high uncertainty, we filtered out clusters for which the estimated standard error was greater than 40% of the estimated value of ρ. To increase computational efficiency, clusters larger than 20 Mb in size were partitioned into smaller subclasses, and estimates of ρ were computed as weighted averages (weighted by the number of informative sites) across subclasses.

1.5.4 Neutral sites

The collection of sites predicted to be free from the influence of natural selection (neutral sites) was derived from a set identified previously[28,34,54]. Briefly, this set was obtained by eliminating from all genomic sites those likely to be under direct natural selection, including (i) exons of annotated protein-coding genes and the 1,000 bp flanking them on either side; (ii) RNA genes from GENCODE v11 and the 1,000 bp flanking them; and (iii) conserved noncoding elements (identified by phastCons) and the 100 bp flanking them. This set was used in both the INSIGHT analysis and the power analysis.

1.5.5 GENCODE annotations

Transcript annotations from GENCODE v15 (ref. [55]) were downloaded from the Sanger Institute's FTP server and used to define eight site classes: CDSs, 5′ UTRs, 3′ UTRs, promoters, introns, lincRNAs, sncRNAs and intergenic (sites not falling within any protein-coding transcription unit). Transcripts annotated with feature type = "CDS" and gene type = "protein coding" were used to define the CDS set for fitCons. For subsequent analysis, we used a slightly more conservative set, obtained by additionally requiring feature type = "gene," gene status = "KNOWN," transcript status = "KNOWN" and the identification of both start and stop codons within the transcript. UTRs were defined from transcripts having feature type = "UTR" and gene type = "protein coding" and were designated as 5′ or 3′. Introns were defined by positions that fell within a protein-coding transcript but outside of the CDS and UTRs.

Promoters were defined as the 1,000 bp immediately upstream of the first (most upstream) transcription start site for each protein-coding gene. A similarly defined alternative set of 100-bp promoter regions was used in assessing differences between cell types (Supplementary Fig. 1.15). lincRNAs were identified by transcripts with feature type = "exon" and gene type = "lincRNA." Similarly, sncRNAs consisted of transcripts with feature type = "exon" and gene type ∈ {"miRNA," "snRNA," "snoRNA"}. Positions in the more inclusive CDS set were removed from all noncoding classes.

1.5.6 *Cis* regulatory elements

Transcription factor binding sites were drawn from a set for 78 transcription factors, based on ChIP-seq data from ENCODE28 downloadable from our UCSC Genome Browser mirror. This set contained roughly 1.4 million binding sites of a mean length of 11 bp, each of which was associated with the cell types in which it was detected. For some tests, we considered only the subset of nucleotide positions inside these transcription factor binding sites that corresponded to motif positions with strong base preferences, defined as those positions at which the consensus allele appeared in at least 90% of all binding sites (according to the inferred motif model). For enhancers, we used the distal regulatory modules described in ref. 38. We downloaded the file enets4.Distal_cell_line.txt from the Gerstein laboratory ENCODE nets and extracted from it a total of 19,005 enhancer-transcript associations, covering 5,834 unique autosomal loci with a mean length of 888 bp, along with the cell types associated with each predicted enhancer. The eQTLs described in ref. 6 were downloaded from the European Bioinformatics Institute's E-GEUV-1 data set. We used the 4 files {EUR373, YRI89}.{exon, gene}.cis.FDR5.best.rs137.txt.gz to identify 6,760 distinct autosomal positions and the associated transcripts, removing all positions overlapping CDSs.

1.5.7 Identifying active elements per cell type

In several analyses, we considered the subset of elements in each annotation class for which we had evidence of activity in a given cell type. To identify the cell types in which transcription factor binding sites and enhancers were active, we used the cell type designations provided in the corresponding annotation files. For other classes of elements, we defined the active elements using a set of GENCODE transcripts and genes that showed significantly elevated levels of RNA transcription in the Caltech RNA-seq data. These were transcripts (or genes) for which the 95% confidence interval of the normalized read count in a given cell type fell within the top one-third of the normalized read counts for transcripts (or genes) across all three cell types (with thresholds of 1.477 for transcripts and 4.966 for genes). Active eQTLs were identified via associated active genes using the GENCODE gene identifier specified for each eQTL. Active promoters, UTRs, CDSs and introns were identified via associated active transcripts. For the comparison between cell types (Supplementary Fig. 1.15), we also used collections of eQTLs and promoters found to be inactive in a given cell type. These were defined in a similar way, by using transcripts and genes falling in the bottom third of the distribution of normalized read counts.

1.5.8 Comparison with other scores

Base-wise scores from the GERP[13] method, the CADD[35] method, and the phastCons[12] and phyloP[15] methods were downloaded from the respective websites (see URLs; file hg19.GERP_scores.tar.gz generated in August 2010 for GERP, file whole_genome_SNVs.tsv.gz downloaded in September 2013 for CADD and the UCSC Genome Browser 46 placental mammal conservation tracks for phastCons and phyloP). CADD scores are specified for each genomic position and each of the three possible variant bases at that position. We took the maximum of these three scores,

which yielded the best performance for the CADD method in our comparisons. We also used RegulomeDB[36] (downloaded in January 2013) to rank SNPs, such as eQTLs, into 1 of 13 categories according to evidence from functional genomic data. Finally, we obtained EnhancerFinder scores[37] for 1,500-bp windows tiled across the genome directly from the authors. We used the general, non-tissue-specific scores and averaged them at positions contained in multiple overlapping windows.

1.5.9 Receiver operating characteristic curves

We used ROC curves to measure the ability of each scoring scheme to discriminate between functional and nonfunctional regulatory elements. For transcription factor binding sites and enhancers, we used the annotations described above as true positives and defined true negatives from our filtered, putatively neutral sites. For eQTLs, our negative set consisted of all 9.8 million variants tested in ref. 6, excluding indels, non-simple variants and positions that showed possible associations at a threshold of nominal $P < 0.05$ (7.6 million SNPs remained). In all three cases, we additionally removed any sites in the positive set from the negative set. A point on a ROC plot indicates the fraction of the annotated genomic positions with scores higher than a given score (true positive rate) versus the fraction of control genomic positions with scores higher than that score (false positive rate). Positions with no scores were ignored when computing fractional coverage.

1.5.10 Integrating fitCons scores across cell types

We generated a series of fitCons scores that integrate functional genomic data across the 3 cell types by using the original 624 fingerprints and altering the rule by which sites are assigned to clusters to reflect information from multiple cell types. Our approach attempts to select a fingerprint for each site that is likely to be most informative about the site's function, while avoiding a bias toward higher scores with an increasing number of cell types. See the Supplementary Note for details.

1.5.11 Share under selection

Assume a partitioning of the genome into K mutually exclusive and exhaustive clusters, $C_1, C_2, \ldots, C_K$, and a corresponding set of fitCons scores, $\rho(C_1), \rho(C_2), \ldots, \rho(C_K)$. Note that the expected number of genomic positions under selection in cluster C_i is given by $\rho(C_i)|C_i|$ because ρ is an estimate of the fraction of sites under selection. For an arbitrary collection of sites S, the expected number of sites in S that are under selection is given by $sel(S) = \sum_i \rho(C_i)|C_i \cap S|$, and the average fitCons score for S is given by $\rho(S) = sel(S)/|S|$. To avoid underestimation of $\rho(S)$, we do not filter out fitCons scores with high uncertainty in these calculations, as we do for other analyses. In addition, to account for possible overestimation of ρ in very large clusters having low fractions of sites under selection, we ran INSIGHT on the intersection of our neutral sites and all noncoding sites in a 'quiescent' chromatin state with no DNase-seq or RNA-seq signal. We then subtracted the estimated value of ρ, denoted ρ_{neut}, from the raw fitCons score to obtain a conservative lower bound, $\rho(S) - \rho_{neut}$, for the fraction of sites under selection in S.

1.5.12 fitConsD and evolutionary turnover

To make the comparison between fitCons and fitConsD as direct as possible, fitConsD scores were computed using the same pipeline we developed for fitCons (Fig. 1.1), except that in step 3 we replaced the INSIGHT model with an evolutionary model that considers sequence divergence between the human, chimpanzee, orangutan and rhesus macaque genomes. fitConsD scores are based on an estimate s_i for the relative evolutionary rate of each cluster C_i in comparison with a neutral model globally estimated for the four-primate phylogeny. This relative rate is then compared with the relative rate s_i^{neut} estimated for the putative neutral regions flanking sites in C_i, and a divergence-based estimate of the fraction of sites under selection in cluster C_i is given by $\rho\ div(C_i) = 1 - s_i/s_i^{neut}$ (Supplementary Note).

1.6 APPENDIX I / SUPPLEMENT TO FIRST PAPER

The following material is not part of the printed version of the paper, nor the full-text online version, however, it is provided by the publisher as an associated resource that is referenced by the online version. In the traditional dissertation format, such material would likely be provided in an Appendix at the end of the document. However, in a papers-format dissertation, guidelines provided by Cornell University request all material from a publication be maintained in the same dissertation chapter. To comply with Cornell University guidelines, relevant appendix material is provided in the following section.

1.6.1 Supplementary Figures

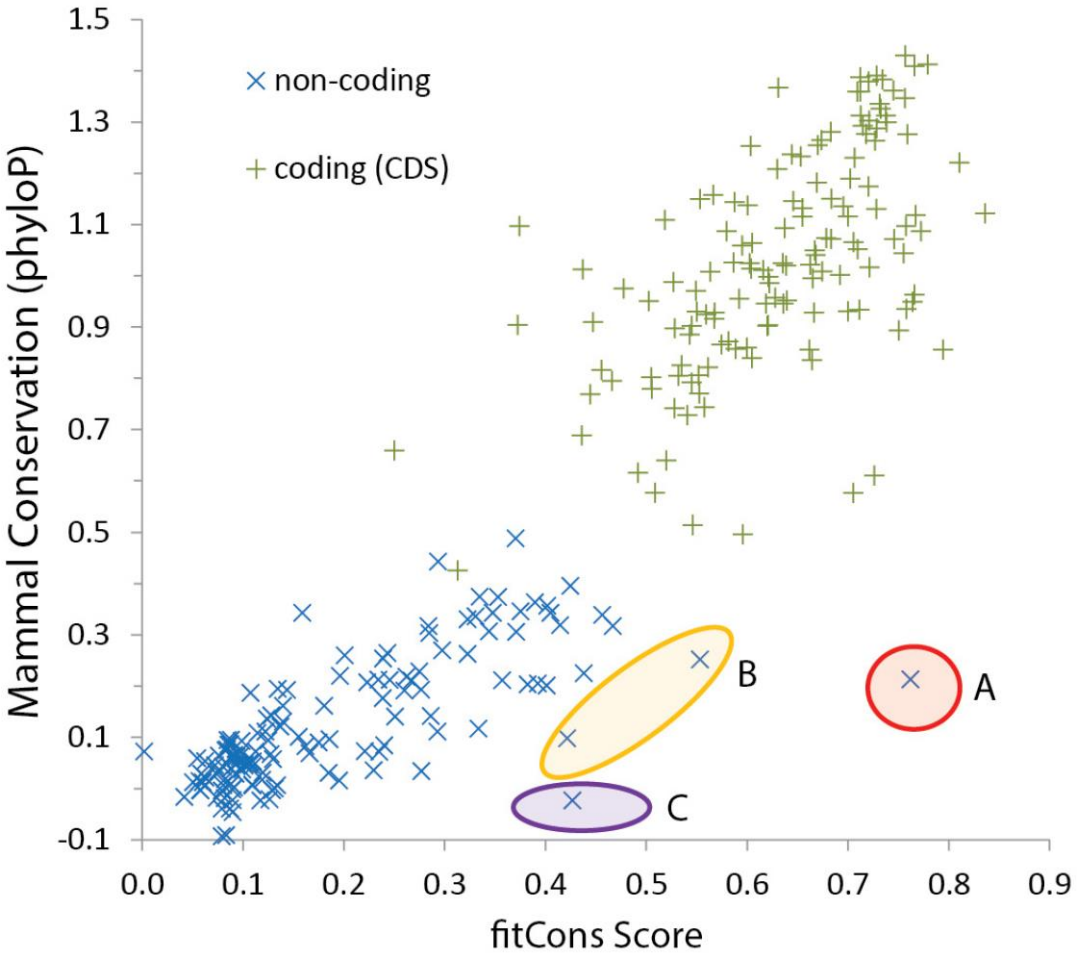


Figure 1.7: Comparison of fitCons scores and phyloP conservation scores. Each of the 624 clusters is represented by a single point, with its x coordinate given by the fitCons score calculated as shown in Figure 1 and its y coordinate given by the mean placental mammalian phyloP score for the associated genomic positions[15]. The clusters naturally fall in two groups, corresponding to coding sequences (CDSs) with higher scores (green crosses) and noncoding sequences with lower scores (blue Xs). Three groups of outliers are shown, representing noncoding clusters with elevated fitCons scores relative to their phyloP scores. Cluster A consists of 1,200 genomic positions in narrow DNase-seq peaks with no RNA-seq signal, yet with chromatin modifications indicating transcription activity. These sites are strongly enriched for ChIP-seq–supported TFBSs and may contain enhancers with weakly expressed eRNAs not detectable from the available RNA-seq data. The two clusters in B contain 92.8 kb of sequence defined by high RNA-seq signals, broad DNase-seq peaks and Pol II binding and are strongly enriched for 3' UTR and ncRNA annotations. Cluster C contains 52.7 kb of sequence with no DNase-seq but some RNA-seq signal, along with insulator associated chromatin modifications. This class is strongly enriched for eQTLs and CTCF-binding sites, suggesting transcriptional silencing activity. Thus, all four of these clusters appear to be rich in regulatory sequences that could plausibly have experienced weak natural selection during most of mammalian evolution but come under stronger selection recently on the human lineage.

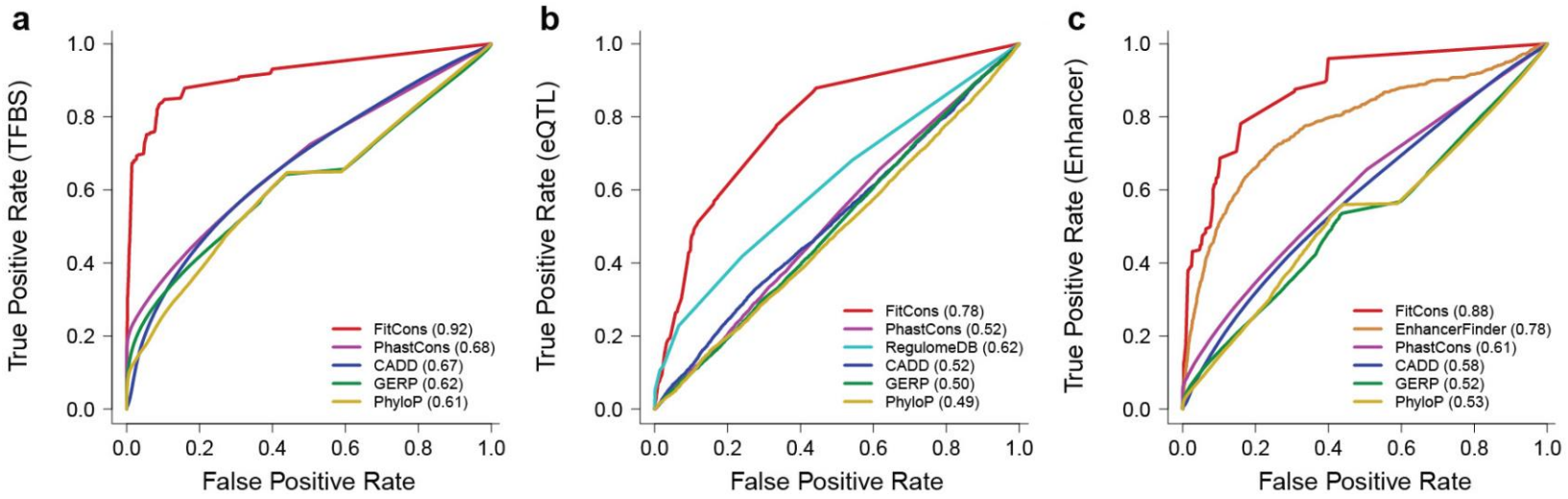


Figure 1.8: Receiver operating characteristic (ROC) curves for cell type–specific regulatory elements. Three types of regulatory elements were considered: (a) transcription factor binding sites (TFBSs), (b) expression QTLs (eQTLs) and (c) enhancers identified by chromatin marks. Separate curves are shown for fitCons, phastCons12, CADD35, GERP13 and phyloP15 scores. In b, a curve is also shown for the RegulomeDB database[36], and in (c) a curve is also shown for EnhancerFinder[37]. True positive rates were estimated by the fraction of nucleotides in annotated elements having scores that exceed a given score threshold, and false positive rates were estimated by the fraction of nucleotides in a matched set of 'negative' elements having scores that exceed the same threshold (see the Online Methods for details). Each curve is generated by varying this threshold across the full range of scores for the corresponding method. In this case, only elements 'active' in the cell type for which the fitCons scores were produced (HUVECs) were considered (Online Methods; see Supplementary Fig. 3 for the results for a pooled set of elements across cell types). AUC values, shown in parentheses, represent areas under the ROC curve and provide an overall measure of predictive power. The apparent performance of RegulomeDB on eQTLs, particularly at low false positive rates, is somewhat influenced by the explicit inclusion of eQTL data in its scoring scheme.

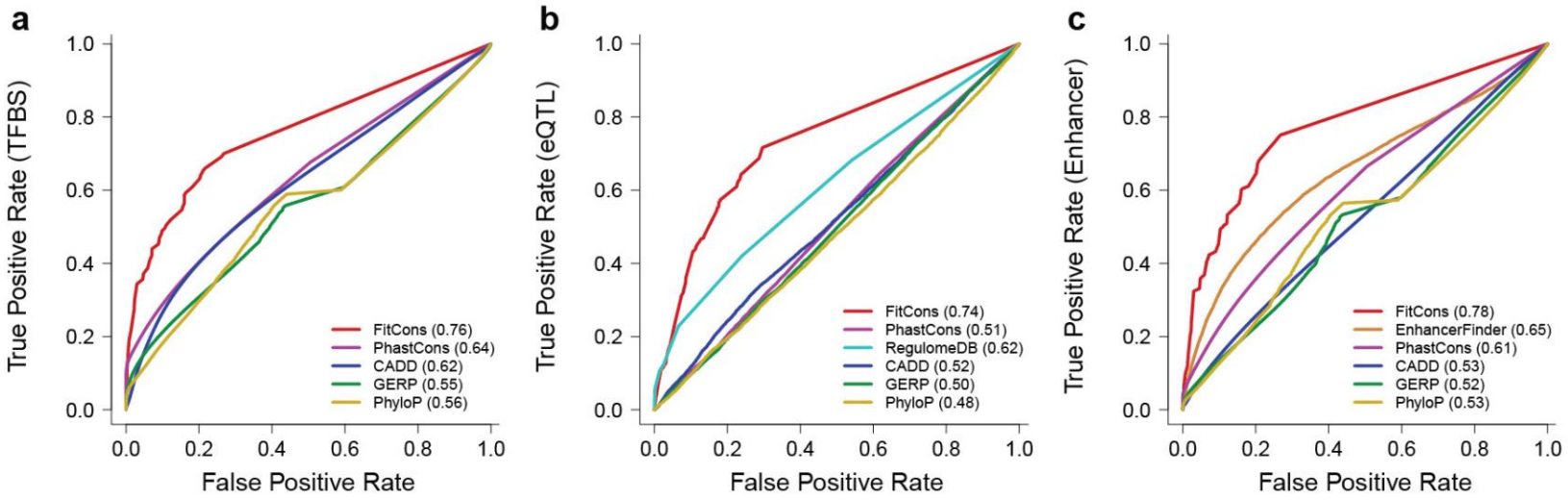


Figure 1.9: Receiver operating characteristic (ROC) curves for regulatory elements pooled across cell types. Three types of regulatory elements were considered: (a) transcription factor binding sites (TFBSs) derived from ENCODE ChIP-seq data for 19 different cell types28, (b) expression QTLs (eQTLs) for lymphoblastoid cells from 462 individuals6 and (c) enhancers identified by chromatin marks in 11 cell types[38]. Separate curves are shown for fitCons, phastCons[12], CADD[35], GERP[13] and phyloP[15] scores. In (b), a curve is also shown for the RegulomeDB database[36], and in (c) a curve is also shown for EnhancerFinder[37]. The fitCons scores used here are computed by aggregating functional information across HUVEC, H1 hESC and GM12878 cells (Online Methods). Note that some regulatory elements might not be active in any of the three cell types. The apparent performance of RegulomeDB on eQTLs, particularly at low false positive rates, is somewhat influenced by the explicit inclusion of eQTL data in its scoring scheme.

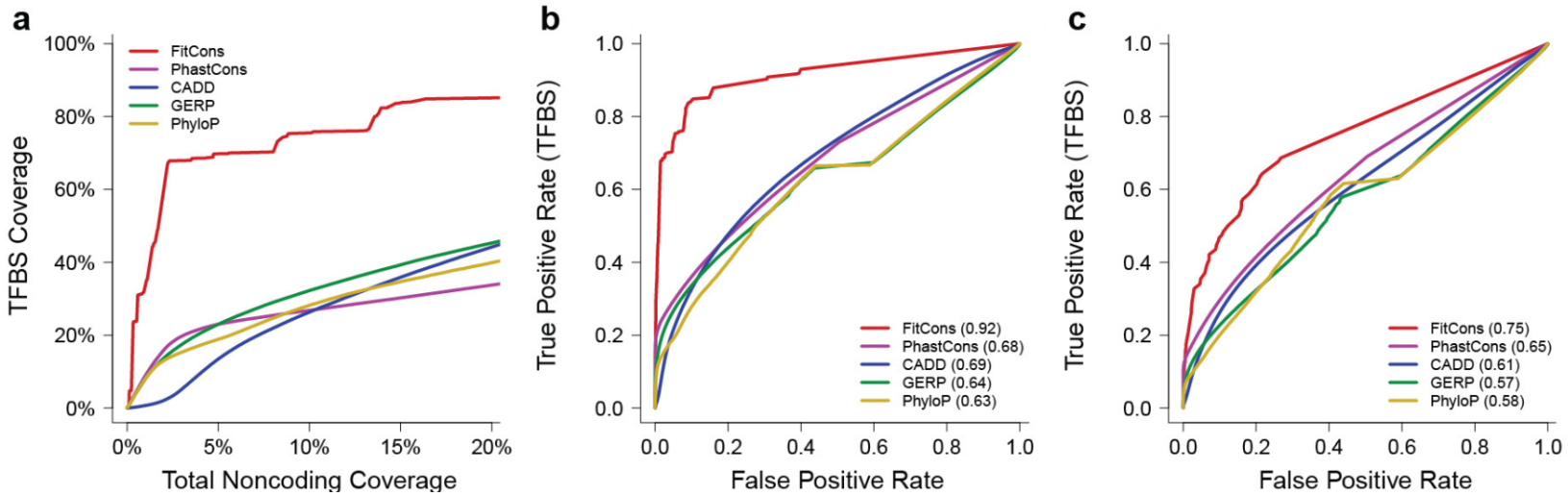


Figure 1.10: ROC and ROC-like curves for high-information-content positions in transcription factor binding sites. These panels parallel previous figures except that, in this case, only positions in ChIP-seq–annotated transcription factor binding sites with strong nucleotide preferences (relative frequency of preferred allele ≥ 90% in motif model) are considered. Shown are (a) coverage as a function of total noncoding coverage (as in Fig. 5a); (b) a receiver operating characteristic (ROC) curve for elements active in HUVECs (as in Supplementary Fig. 2a); and (c) a ROC curve based on elements active in various cell types and integrated fitCons scores (as in Supplementary Fig. 3a). These curves are highly similar to the ones based on whole binding sites, despite known correlations between natural selection and information content for at least some transcription factors[15,28], apparently because these correlations tend to be fairly weak and transcription factor specific and generally occur below the prediction thresholds of interest.

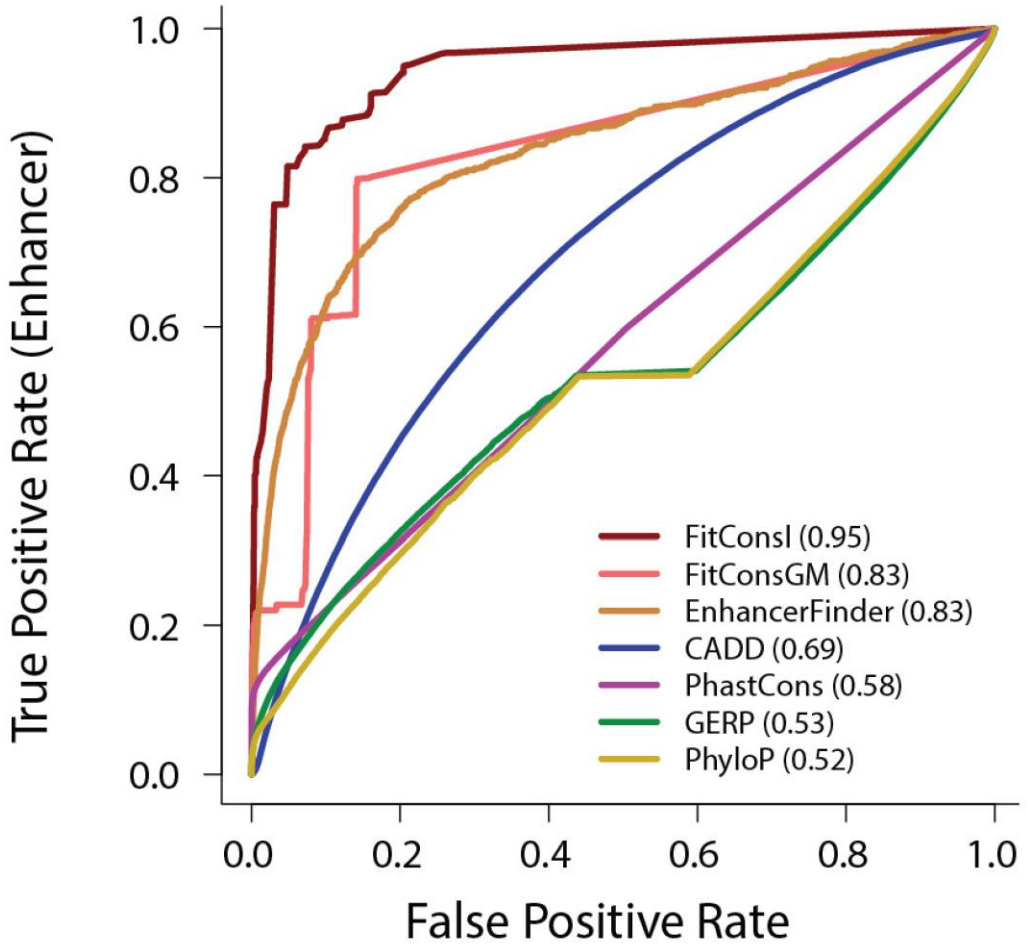


Figure 1.11: Receiver operating characteristic (ROC) curves for alternative enhancer. divergent transcription initiation, as measured by a variant of GRO-seq that enriches for 5'-7meGTP-capped RNAs[39]. The tested enhancers were identified by starting with the 'unstable/unstable' (UU) pairs of divergent transcription start sites from ref. 39 and eliminating those that fell within 2 kb of a known gene. Each enhancer was assumed to consist of a 200-bp interval centered on the midpoint between the paired transcription start sites. Shown are curves for both cell type–integrated (FitConsI) and GM12878-specific (FitConsGM) fitCons scores, as well as for EnhancerFinder[37], CADD[35], phastCons[12], GERP[13] and phyloP[15]. The coarse, stair-step appearance of the FitConsGM curve reflects a lack of diversity in the functional genomic fingerprints coinciding with these enhancers, and the improvement in the FitConsI curve suggests a gain in power from considering overlapping enhancers in other cell types. Notice that EnhancerFinder and CADD perform fairly well on this set, but the conservation-based methods perform poorly.

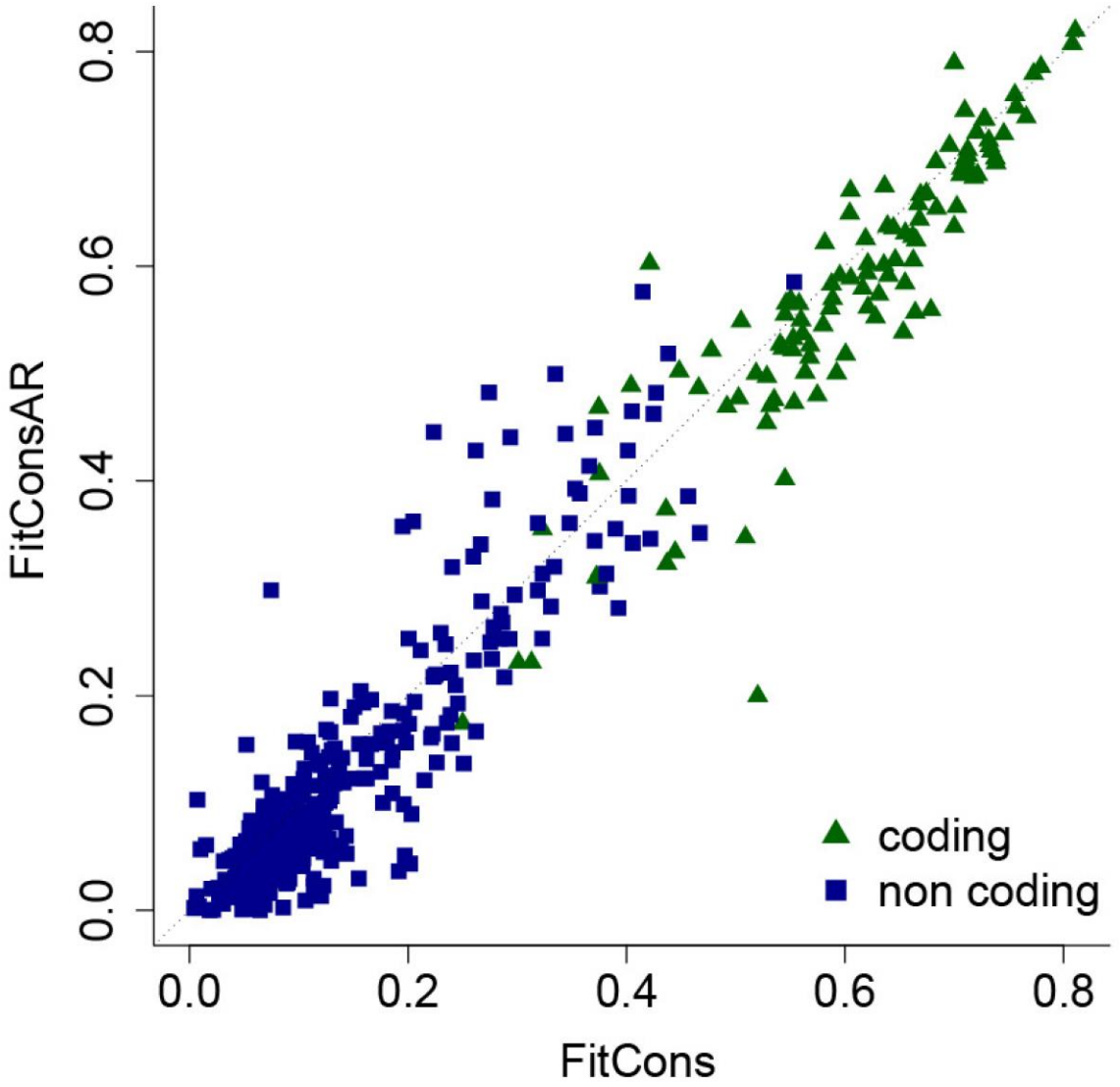


Figure 1.12: Comparison of original fitCons scores (FitCons) with an alternative set of scores based on ancestral repeats as neutral sites. Each point represents a particular functional genomic class. The two sets of scores are highly correlated overall (R^2 = 0.95), suggesting that they are not highly sensitive to the choice of neutral sites. Surprisingly, however, the scores based on ARs are slightly reduced overall (genomic average of 0.058 versus 0.075), apparently owing to reduced estimates of neutral divergence rates for ARs. Notice that this trend is the opposite of what would be expected if the ARs were under less constraint than our more inclusive set of putatively neutral sites, as one might surmise would be true. We speculate that it may be a consequence of unusual properties of transposable elements, such as AT richness, hypermethylation or exapted functional elements. The ARs used for this analysis consisted of families of RepeatMasker-identified repeats having an average divergence from the consensus of >15%, excluding simple sequence repeats, microsatellites, rRNAs, tRNAs and other potentially problematic families (871 Mb of sequence in total).

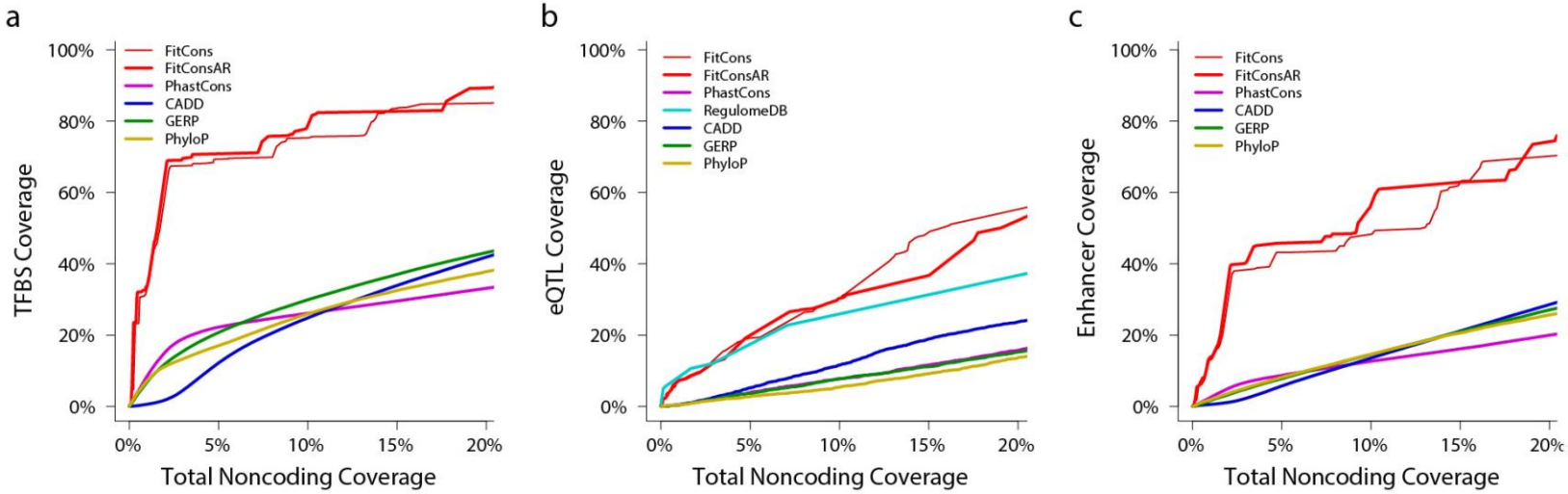


Figure 1.13: Coverage of regulatory elements as a function of total noncoding coverage for fitCons scores based on ancestral repeats. As in Figure 5, coverage of each type of element is shown as the score threshold is adjusted to alter the total coverage of noncoding sequences in the genome. FitCons scores based on ancestral repeats (FitConsAR) are compared with ordinary fitCons scores (FitCons) and scores from phastCons[12], CADD[35], GERP[13], phyloP[15] and RegulomeDB[36]. Notice that the FitCons and FitConsAR scores behave almost identically at low levels of coverage and show only modest differences at higher levels of coverage. See Supplementary Figure 6 for details regarding ancestral repeats.

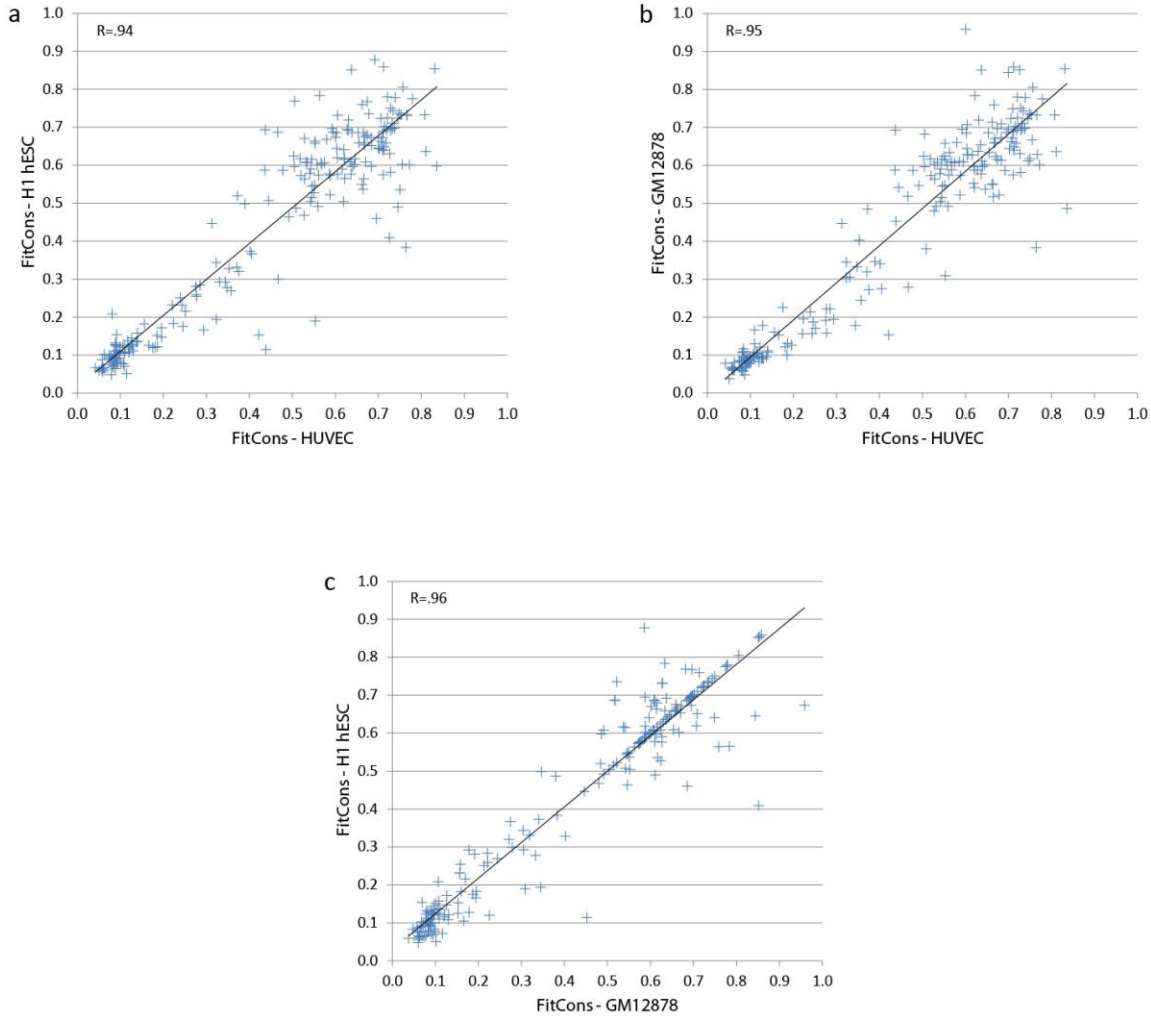


Figure 1.14: FitCons scores for the same functional fingerprint in differing cell types are strongly correlated. FitCons scores for all functional classes for (a) HUVECs versus H1 hESCs, (b) HUVECs versus GM12878 cells, and (c) GM12878 cells versus H1 hESCs. Although the individual positions assigned to each class vary widely according to cell type, the fitCons scores remain relatively constant, with Pearson correlations ≥ 0.93 and Spearman correlations ≥ 0.87 between pairs of cell types.

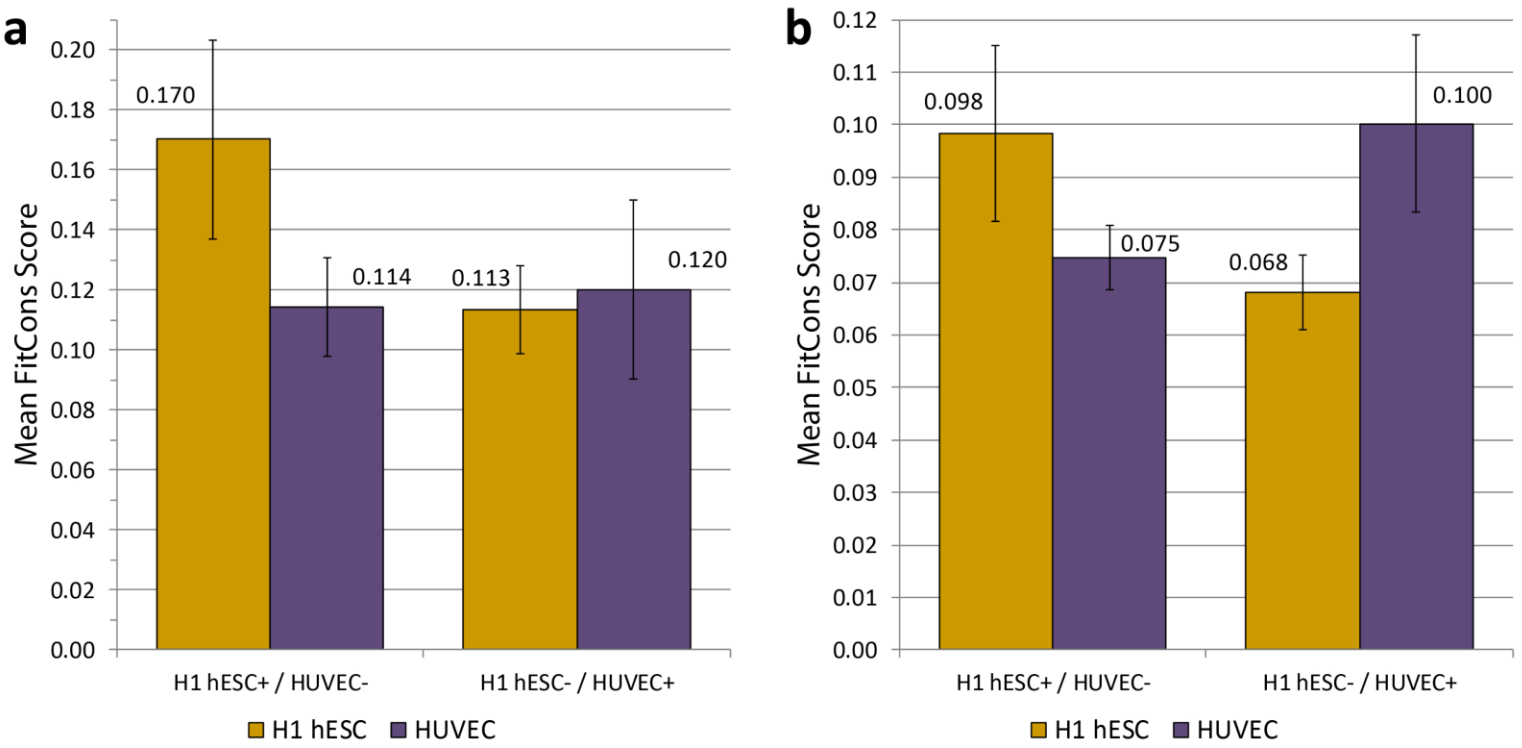


Figure 1.15: FitCons scores reflect cell type–specific activity. Mean fitCons score for (a) 100-bp promoters and (b) eQTLs that are active in one cell type and inactive in another, based on RNA-seq data for the associated gene (Online Methods). Error bars represent the standard errors of the aggregated fitCons scores (Online Methods). FitCons scores computed using functional genomic data from H1 hESCs (orange bars) for elements active in H1 hESCs and inactive in HUVECs (H1 hESC+/HUVEC–) are significantly higher than those for elements inactive in H1 hESCs and active in HUVECs (H1 hESC–/HUVEC+). The opposite pattern is observed for fitCons scores computed using functional genomic data from HUVECs (purple bars).

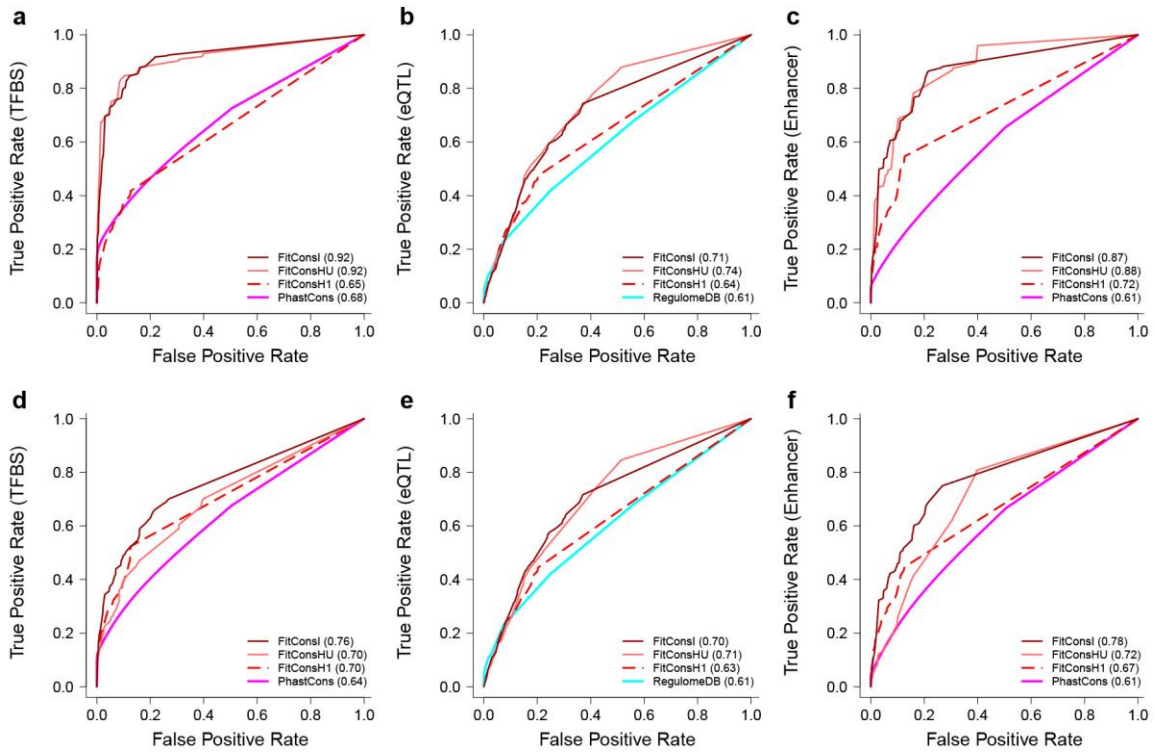


Figure 1.16: Receiver operating characteristic (ROC) curves comparing integrated fitCons scores with cell type–specific fitCons scores. The top row shows the predictive performance of fitCons scores for elements 'active' in the HUVEC cell type: (a) TFBSs, (b) eQTLs and (c) enhancers. Three versions of the fitCons score are shown: cell type–specific scores based on HUVECs (FitConsHU) and H1 hESCs (FitConsH) and scores based on integrated data from all three cell types (FitConsI). Notice that the FitConsI scores perform as well as those based on the 'active' cell type (FitConsHU), whereas those based on a different cell type (FitConsH1) perform substantially worse. The bottom row shows the same fitCons scores applied to elements aggregated from a broad range of cell types: (d) TFBSs, (e) eQTLs and (f) enhancers. In this case, FitConsI outperforms both sets of cell type–specific scores. Thus, the integrated scores (FitConsI) appear to improve performance in a cell type–general setting without much cost in the cell type–specific setting.

Table 1.1: FitCons Site Clusters.

Cell type	Number of clusters			Coverage[d]	Median size[e]
	Non-empty [a]	Viable [b]	Confident [c]		
HUVEC	560	502	295	99.5%	120 kb
H1 hESC	557	498	298	99.1%	121 kb
GM12878	556	496	287	99.4%	144 kb
Integrated	494	466	283	98.9%	198 kb

[a]Number of non-empty clusters in partitioning.

[b]Number of non-empty clusters after applying INSIGHT filters.

[c]Number of clusters for which the ratio between the estimated standard error for ρ and the estimated value of ρ is at most 0.4.

[d]Percentage of bases in confident clusters out of 2,881,033,286 bp in human autosomes.

[e]Median size (in kilobases) of confident clusters.

Table 1.2: Share Under Selection for Various Annotation Classes

Class	Size (Mbp)	% genome	% sites under sel.[a]	% sel. sites[b]	Adj. % sel. sites[c]
genome	2,881.0	100.0%	7.5%±0.1%	100.0%	100.0%
coding	30.7	1.1%	63.3%±0.4%	9.0%	15.3%
3' UTR	24.7	0.9%	19.5%±0.3%	2.2%	3.3%
5' UTR	4.1	0.1%	13.3%±0.5%	0.3%	0.3%
promoter	21.5	0.7%	9.0%±0.3%	0.9%	1.0%
ncRNA[d]	8.0	0.3%	8.1%±0.1%	0.3%	0.3%
intron	1,008.8	35.4%	7.5%±0.1%	35.2%	35.1%
intergenic	1,768.2	61.5%	6.3%±0.1%	51.7%	44.1%

[a]Fraction of sites under selection in a class computed using HUVEC fitCons scores (see Methods).

[b]Fraction of total number of sites under selection in the genome that is estimated to fall in class.

[c]Fraction of sites under selection estimated to fall in class, corrected by subtracting $\rho_{\text{neut}} = 3.3\%$ from the raw estimate (see Methods).

[d]Union of lincRNA set and sncRNA set.

Table 1.3: Sources of Functional Genomic Data.

1.6.3 Supplementary Note

1.6.3.1 Partitioning Genome Based on RNA-seq Data

The coarse-grained, discrete DNase-seq and histone modification data (with broad vs. narrow peaks and the 25 ChromHMM states) naturally partitioned the genome into a small number of classes. Using the RNA-seq data for partitioning, however, required developing a framework that would allow us to determine how informative a given partition is on the distribution of sites under selection in the genome. To address this problem, we searched exhaustively for a maximally informative partitioning, using the measure of mutual information as our objective function. This exhaustive search was carried out by dividing the range of continuous values (normalized read depth in the case of RNA-seq) into a discrete set of intervals, and assessing the fraction of sites under selection in each interval using INSIGHT. This approach provides a general framework for using fitCons scores computed by INSIGHT to refine a given clustering scheme (see backward arrow from C to B in Fig. 1.1), and we anticipate that it will be useful in parsing other complex data sets. The three sections below describe the information theoretic concepts used in our approach, the implementation details of the exhaustive search, and the results for the RNA-seq data for the three cell types.

1.6.3.2 Mutual Information and Conditional Entropy

Let X be a binary variable indicating whether or not a genomic position is under selection; that is, if a mutation at that site will influence fitness then $X = 1$ and otherwise $X = 0$. In addition, let Y_C indicate the cluster to which the same position is assigned in a given partitioning $C(Y_C \in C = \{C_1, \ldots C_K\})$. Assuming that sites are selected uniformly at random from the genome and that $\rho(C_i)$ denotes the fraction of sites under selection in cluster C_i, the joint probability distribution of X and Y_C is given by:

$$P(X = 1, Y_C = C_i) = \frac{|C_i|}{\sum_j |C_j|}\rho(C_i) \quad (1)$$

For notational simplicity below, let $\rho = \frac{\sum_i |C_i|\rho(C_i)}{\sum_i |C_i|}$, the fraction of sites under selection in the genome. The mutual information of X and Y_C is given by the following expression[56] :

$$I(X; Y_C) = \sum_{i=1}^{k}\sum_{x=0}^{1} P(X = x, Y_C = C_i)\log\left(\frac{P(X = x, Y_C = C_i)}{P(X = x)P(Y_C = C_i)}\right) \quad (2)$$

$$= \sum_{i=1}^{k} \frac{|C_i|}{\sum_j |C_j|}\left(\rho(C_i)\log\left(\frac{\rho(C_i)}{\rho}\right) + \big(1 - \rho(C_i)\big)\log\left(\frac{\big(1 - \rho(C_i)\big)}{(1 - \rho)}\right)\right) \quad (3)$$

$$= \sum_{i=1}^{k} \frac{|C_i|}{\sum_j |C_j|}\big(\rho(C_i)\log\big(\rho(C_i)\big) + \big(1 - \rho(C_i)\big)\log\big(1 - \rho(C_i)\big)\big) + H(X) \quad (4)$$

Note that $H(X) = -(\rho\log(\rho) + (1 - \rho)\log(1 - \rho))$, the entropy of X, does not depend on the partitioning C, and the remaining terms on the right-hand side of the last equation are equal to $-H(X|Y_C)$, where $H(X|Y_C)$ denotes the conditional entropy of X given Y_C. Thys maximizing the mutual information of $I(X; Y_C)$ is the same as minimizing the conditional entropy $H(X|Y_C)$.

1.6.3.3 Implementation

Our method for partitioning the genome into K read-depth bins (for a given K) is based on an exhaustive search of all ordered K-partitions, C, to find the one that results in the largest mutual information $I(X; Y_C)$. To make the exhaustive search tractable, we apply it to discretized partition boundaries using the procedure outlined below:

1. Divide the continuous range of values (normalized RNA-seq read depth in our case) into N discrete intervals, $I_1, \ldots, I_N$, such that intervals are

of comparable size and large enough to produce confident estimates of ρ using INSIGHT.

2. Run INSIGHT on the collection of sites corresponding to each interval I_i to obtain an estimate, of the fraction of sites under selection in I_i.
3. For each of the $\binom{N}{2}$ ordered pairs $1 \leq i \leq N$, denoted by $I_{i,j}$, the union of all I_k, such that $k \in [i,j]$, and estimate $\rho(I_{i,j})$ using the weighted average $\rho(I_{i,j}) = \frac{\sum_{k\in[i,j]}|I_k|\rho(I_k)}{\sum_{k\in[i,j]}|I_k|}$.
4. For each of the $\binom{N-1}{K-1}$ discretized K-partitions, $C = \{C_1, \ldots, C_K\}$, defined by $K+1$ interval boundaries,

$$0 = i_1 \leq i_2 < i_3 < \cdots < i_K < i_{K+1} = N$$

retrieve for each cluster $C_k = I_{(i_k+1),i_{k+1}}$, an estimate of $\rho(C_k)$ from the estimates pre-computes above, and use it to compute the mutual information $I(X;Y_C)$, using the expression in equation (4), above.

5. Choose the K-partiotion with the highest mutual information.

Applying this procedure with increasing values of K should result in an increase in the resulting mutual information, but a decrease in the size of clusters.

1.6.3.4 Application to RNA-seq data

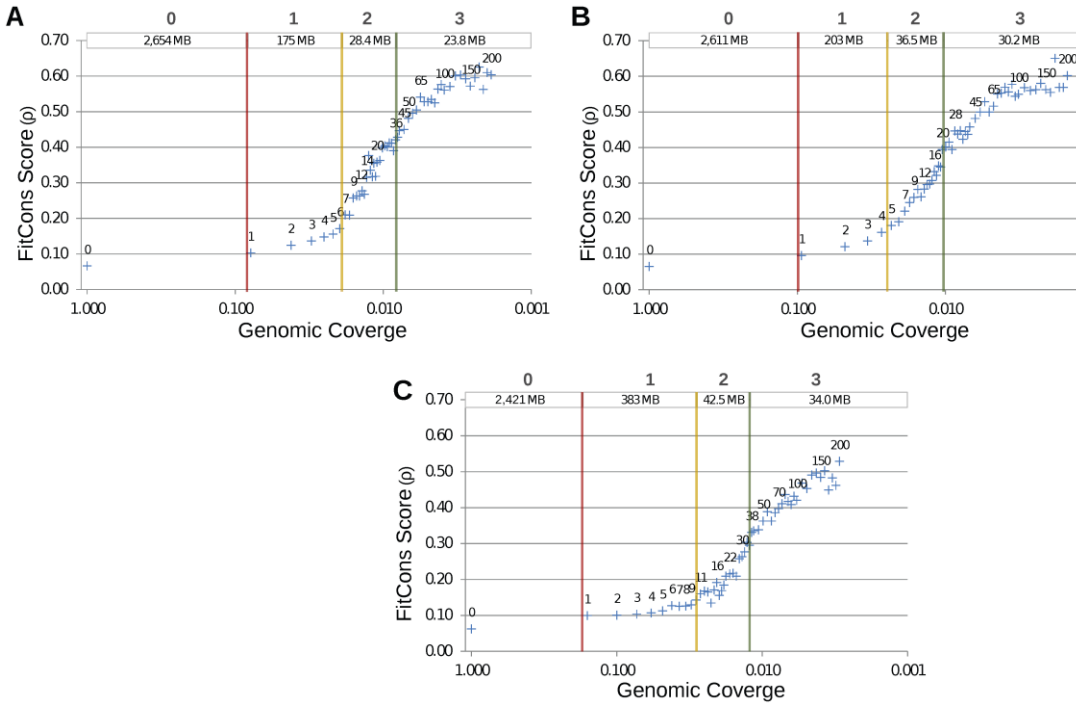
We applied the procedure described above separately to the RNA-seq data of each of the three cell types. For each cell type, we divided the range of normalized read depth (reads per million; RPM) into N=53 intervals by taking increments of 1 RPM between 0 and 20, increments of 2 between 20 and 40, increments of 5 between 40 and 100, increments of 10 between 100 and 200, and allocating a single interval for

RMP>200. We computed $\rho(I_i)$ for each of the 53 intervals (step 2 in the procedure described above), and used it to compute estimates of ρ for each of the possible $\binom{53}{2}$ discrete bins (step 3). Then we executed the exhaustive search (steps 4–5) for K = 2, 3, 4, 5 (see table below). While the mutual information $I(X;Y_C)$ kept increasing as we increased K, partitioning into more than 4 bins resulted in small bins (less than 30 Mb) for intermediate read depths, which we did not expect to be very informative. We thus chose K = 4 for our final partitioning. Note that this partitioning results in one boundary at RPM=1, another boundary near the deflection point for$\rho(I_i)$, and a third boundary around the middle of the dynamic range of ρ estimates (see figure below).

Partitioning Based on RNA-seq Read Depth

cell type	$\mathcal{C}$ [a] ; $I(X;Y_C)$ [b] (K=2)	$\mathcal{C}$; $I(X;Y_C)$ (K=3)	$\mathcal{C}$; $I(X;Y_C)$ (K=4) [c]	$\mathcal{C}$; $I(X;Y_C)$ (K=5)
HUVEC	$\{7\}$; 0.0117	$\{1, 14\}$; 0.0133	$\{1, 7, 36\}$; 0.0138	$\{1, 6, 14, 50\}$; 0.0140
H1 hESC	$\{7\}$; 0.0116	$\{2, 15\}$; 0.0132	$\{1, 5, 20\}$; 0.0139	$\{1, 3, 7, 20\}$; 0.0141
GM12878	$\{20\}$; 0.0093	$\{1, 28\}$; 0.0120	$\{1, 11, 38\}$; 0.0124	$\{1, 10, 28, 75\}$; 0.0126

[a] Partition $\mathcal{C}$ is indicated by $K-1$ thresholds of RPM from our pre-determined set of 52 boundary points.
[b] $I(X;Y_C)$ is the mutual information of the partitioning $\mathcal{C}$ and the distribution of sites under selection, as defined in Equation 4.
[c] $K = 4$ was used for the partitioning in fitCons.



Partitioning the genome into K=4 bins according to normalized read depth in RNA-seq experiments for HUVEC (panel A), H1 hESC (panel B), and GM12878 (panel C). Each point in the scatter plot represents one of N=53 (atomic) intervalsI_i, plotting $\rho(I_i)$ as a function of the fraction of the genome covered by the union of intervals $I_i, I_{i+1}, \ldots, I_N$. The label next to each point corresponds to the lower boundary (in RPM) of that interval. Note that $\rho(I_i)$ typically increases with i, indicating a higher concentration of sites under selection in highly transcribed sequences. The boundaries between the four resulting classes (0-3) are indicated by vertical lines, with labels (top) representing the class designation and the number of position in each class.

1.6.3.5 Controlling for Model Misspecification in INSIGHT

The INSIGHT model makes several simplifying assumptions that could potentially influence its estimates of ρ. While these assumptions should not generally bias estimates in any particular direction, the fact that ρ is restricted to be positive might lead to a slight bias when estimating ρ for site clusters that have a near zero fraction of sites under selection. This slight bias might have a nonnegligible influence on our estimate of the fraction of nucleotides under selection (7.5%), because this

estimate is obtained by taking a weighted average of estimates of ρ across all clusters, and the terms dominating this average belong to large clusters with very low fractions of sites under selection. To estimate the potential effect of this bias, we ran INSIGHT on the collection of sites that belong to our putatively neutral set and have a null functional fingerprint, i.e., DNase-seq and RNA-seq classes 0, 'quiescent' chromatin state, and non-CDS. Our expectation is that INSIGHT should infer $\rho = 0$ for this collection of sites, because it is depleted in functional sites, and more importantly, the putatively neutral sites are used by INSIGHT to define the neutral model. Estimating ρ for this very large collection of sites (790 Mb) was done by dividing it into sub-clusters smaller than 20Mb, running INSIGHT on each sub-cluster and taking the weighted average of the resulting estimates (same approach was used in our main pipeline for large site clusters). The resulting estimate of $\rho_{neut} = 0.033$ was then subtracted from the estimates of each site clusters to obtain a conservative lower bound for ρ for that cluster.

1.6.3.6 Differences Between Cell Types

Our main analysis focuses on HUVEC, but we also generated fitCons scores for H1 hESC and GM12878. To compare the scores for different cell types, we began by examining the 624 functional genomic classes across the three cell types, in terms of both the genomic positions assigned to each class, and the fitCons scores estimated for those positions. (Note that our partitioning scheme ensures that the same 624 class definitions are used for each cell type.) Approximately 30% of genomic positions had a null functional fingerprint in all three cell types. In the remainder, we found that genomic positions assigned to each class differed substantially across cell types, with fewer than 4.5% of positions being assigned to the same functional class across all three cell types, and more than a third being assigned to different functional classes in all three cell types. Despite their association with different genomic positions,

however, equivalently defined clusters exhibited highly similar fitCons scores across cell types (Pearson correlation ≥ 0.93 for all pairs; Supplementary Fig. 1.14). Thus, while the patterns of activity differ substantially across cell types, the evolutionary signatures associated with genomic positions that display particular patterns of activity are remarkably consistent across cell types.

To examine the degree to which the scores convey cell-type-specific information, we next considered fitCons scores for elements that are active in one cell type an inactive in another. In particular, we examined subsets of eQTL and proximal promoters (within 100bp of the annotated transcription start site) that appear to be active in H1 hESC but inactive in HUVEC (H1 hESC+/HUVEC−) or inactive in H1 hESC and active in HUVEC (H1 hESC−/HUVEC+) based on RNA-seq data for the same cell types (see Methods). For each of these groups of elements, we compared mean fitCons scores computed for each of the two cell types (H1 hESC and HUVEC). We found that, based on the scores computed for each cell type, the active elements in that cell type had significantly higher scores than the inactive elements (compare the two gold bars and the two purple bars in each panel in Supplementary Fig. 1.15). In addition, the same sets of functional elements have significantly higher fitCons scores for the cell type in which they are active than for the one in which they are inactive (compare adjacent gold and purple bars in Supplementary Fig. 1.15). Similar patterns were observed for comparisons involving GM12878 (results not shown). These findings demonstrate that, while the fitCons scores for all cell types are based on the same polymorphism and divergence data, they nevertheless convey cell-type-specific information through the use of cell-type-specific functional data for clustering.

1.6.3.7 Integrating fitCons Scores across Cell Types

Despite the advantages of the cell-type-specific scores, it is sometimes desirable to have single set of scores that integrate information from multiple cell

types. The main challenge in generating fitCons scores that integrate functional genomic data across cell types, within the context of our simple partitioning scheme, is avoiding a combinatorial explosion in the number of functional genomic clusters considered. We addressed this problem by fixing the partitioning scheme to the original 624 fingerprints, but altering the rule by which nucleotide sites are assigned to clusters to reflect information from multiple cell types. In particular, we attempted to select, for each nucleotide site, the single cluster from all clusters to which that site was assigned across cell types that was likely to be most informative about the site's function. Toward this end, we computed a cell-type aggregated estimate of ρ for each of the 624 classes by running INSIGHT on the collection of all sites associated with that class in any of the three cell types. Note that, unlike in the standard fitCons pipeline (see Fig. 1.1), these collections of sites overlap with one another. We then partitioned the sites into non-overlapping clusters by choosing, for each genomic position, the cluster (out of the three) that had the highest cell-type aggregated ρ. Finally, we executed INSIGHT once more on each of these disjoint clusters to obtain cell-type integrated fitCons scores. We settled on this strategy after observing that a simpler two-pass approach—in which we assigned each position to the single class that maximized its score across cell types and then re-estimated the scores accordingly—tended to cause the "null" class to grow as the number of cell types increased, which decreased the dynamic range of the scores. Our more complex strategy avoided this problem.

Note that this approach produces scores that have nearly equal predictive power for elements active in all three cell types, and much better power than the cell-type-specific scores have when they are applied to a mismatched cell type (Supplementary Fig. 1.16). The integrated scores are publicly available along with the cell-type-specific scores in our genome browser tracks (http://genome-mirror.cshl.edu,

hg19 assembly). We recommend the use of these scores when scores for a matching or similar cell type to the one of interest are not available.

1.6.3.8 FitConsD

The purpose of the fitConsD scores is to provide a measure of natural selection over longer evolutionary timescales, namely, since the divergence of human, chimpanzee, orangutan, and rhesus macaque. This measure is designed to be methodologically as close as possible to that used for the fitCons scores. We thus used the same pipeline described in Fig. 1.1, except that in step C we replaced the INSIGHT model with an evolutionary model that considers sequence divergence between the four primate genomes. This procedure is described in detail below.

We downloaded the multiple genome alignment for 46 placental mammals from the UCSC Genome Browser (http://genome.ucsc.edu), and extracted from it the subalignment for the four primates. In each of the three non-human genomes, we filtered out nonsyntenic regions and positions with genotype quality below 20. Additionally, we masked out sites filtered in the INSIGHT analysis to eliminate repetitive sequences, recent duplications, and CpG sites (as filtered in our INSIGHT analysis). We assumed a fixed branch-weighted phylogeny T for the four-species tree, which was obtained by fitting a phylogenetic substitution model to fourfold degenerate sites in coding sequences, and we used the partitioning of the genome into 624 clusters $\{C_i\}$ defined using functional data from the HUVEC cell line.

With these preparations, we estimated a divergence-based fraction of sites under selection, $\rho_{div}(C_i)$ for each clusterC_i, as follows. First, we created a pseudo-alignment consisting of the columns from the original four-species alignment that correspond to positions in Ci. We then used the *phyloFit* procedure from RPHAST[57] to estimate a maximum-likelihood scaling factor s_i for the tree T for this pseudoalignment. This scaling factor s_i is an estimate of the relative evolutionary rate

in cluster C_i compared with the pre-estimated neutral model, but it does not yet consider variation in the neutral substitution rate along the genome. Therefore, we additionally computed similar scale factors for 10 kb blocks of neutral sites across the genome, using the same neutral sites and windowing scheme as used by INSIGHT. Specifically, for each 10 kb window w, we computed a maximum-likelihood neutral scaling factor s_w^{neut} for T. We then defined the neutral scale factor s_i^{neut} for a cluster C_i as the weighted average of neutral scale factors $\{s_w^{neut}\}$ in the associated neutral blocks (i.e., the average is weighted by the size of the intersection of cluster C_i and each window w). Now the relative rate of substitution in C_i compared to the expectation under neutrality could be computed as, s_i/s_i^{neut}. Under the assumption that negative selection dominates[58], an estimate of the fraction of sites under selection in cluster C_i is therefore given by $\rho_{div}(C_i) = 1 - s_i/s_i^{neut}$, and this is the fitConsD score associated with all sites inC_i.

1.6.3.9 Evolutionary vs. Biochemical Measures of "Function"

Following the publications by the ENCODE Consortium in 2012, there has been a great deal of discussion in the scientific literature, the scientific press, and social media about the discordance between evolution-based estimates of the SUS and estimates of the "functional" content of the genome based on high-throughput measures of biochemical activity, which have been reported to be as high as 80%[3,52]. For various reasons, the ENCODE-based claims do appear to require a rather generous definition of "function"[16–19]. Nevertheless, it is worth emphasizing that the question of the functional content of the genome is inevitably dependent on how function is defined.

Consider two possible definitions of "functional" DNA sequences: sequences that produce a phenotype either (1) when mutated (by point mutations), or (2) when deleted. Under the first definition, genomic positions such as fourfold degenerate sites

in coding regions or degenerate positions in TFBSs will generally not be functional, whereas under the second definition they will be functional, because their presence is required to maintain the functional coherence of a larger element (they are both examples of "spacer" elements). Other examples of functional sequences whose function does not depend on the precise identity of each nucleotide at each position include sequences separating binding sites for interacting TFs, sequences in short introns, and sequences that maintain the spacing properties of *cis*-regulatory elements relative to target genes.

Importantly, most estimates of the SUS, including ours, have made use of definition (1), whereas measures of biochemical activity are more consistent with definition (2) in some respects (although not all spacer elements will be biochemically active). In our view, it is unlikely that this distinction can account for the difference between estimated genomic fractions of ~80% and ~5%. Nevertheless, it is worth bearing in mind that our estimate of the SUS and those from comparative genomics are based on a fairly restrictive definition of function. Indeed, our methods indicate that the SUS in annotated coding regions is only about 60%, a fraction that would undoubtedly rise under definition (2).

REFERENCES

1. Mardis, E. R. A decade's perspective on DNA sequencing technology. Nature 470, 198–203 (2011).
2. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. Nat. Methods 5, 19–21 (2008).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
4. Shen, Y. et al. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120 (2012).
5. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83–90 (2012).
6. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013).
7. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. 12, 628–640 (2011).
8. Mayor, C. et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16, 1046–1047 (2000).
9. Margulies, E. H., Blanchette, M., Program, N. C. S., Haussler, D. & Green, E. D. Identification and Characterization of Multi-Species Conserved Sequences. Genome Res. 13, 2507–2518 (2003).
10. Boffelli, D. et al. Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. Science 299, 1391–1394 (2003).
11. Ovcharenko, I., Boffelli, D. & Loots, G. G. eShadow: A Tool for Comparing Closely Related Sequences. Genome Res. 14, 1191–1198 (2004).
12. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005).
13. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15, 901–913 (2005).
14. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of Sequence Conservation at Nucleotide Resolution. PLOS Comput. Biol. 3, e254 (2007).
15. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121 (2010).
16. Graur, D. et al. On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE. Genome Biol. Evol. 5, 578–590 (2013).

17. Niu, D.-K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? Biochem. Biophys. Res. Commun. 430, 1340–1343 (2013).

18. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. Proc. Natl. Acad. Sci. 110, 5294–5300 (2013).

19. Eddy, S. R. The ENCODE project: Missteps overshadowing a success. Curr. Biol. 23, R259–R261 (2013).

20. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652–654 (1991).

21. Fay, J. C., Wyckoff, G. J. & Wu, C.-I. Positive and Negative Selection on the Human Genome. Genetics 158, 1227–1234 (2001).

22. Andolfatto, P. Adaptive evolution of non-coding DNA in Drosophila. Nature 437, 1149–1152 (2005).

23. Eyre-Walker, A., Woolfit, M. & Phelps, T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. Genetics 173, 891–900 (2006).

24. Boyko, A. R. et al. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. PLOS Genet. 4, e1000083 (2008).

25. Wilson, D. J., Hernandez, R. D., Andolfatto, P. & Przeworski, M. A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences. PLOS Genet. 7, e1002395 (2011).

26. Ward, L. D. & Kellis, M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science 337, 1675–1678 (2012).

27. Khurana, E. et al. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science 342, 1235587 (2013).

28. Arbiza, L. et al. Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. 45, 723–729 (2013).

29. Narlikar, L. et al. Genome-wide discovery of human heart enhancers. Genome Res. 20, 381–392 (2010).

30. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. Nat. Methods 11, 294–296 (2014).

31. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 28, 817–825 (2010).

32. Hoffman, M. M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods 9, 473–476 (2012).

33. Hoffman, M. M. et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841 (2013).

34. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. Mol. Biol. Evol. 30, 1159–1171 (2013).

35. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).

36. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–1797 (2012).

37. Erwin, G. D. et al. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. PLOS Comput. Biol. 10, e1003677 (2014).

38. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–100 (2012).

39. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. 46, 1311–1320 (2014).

40. Chinwalla, A. T. et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002).

41. Cooper, G. M. et al. Characterization of Evolutionary Rates and Constraints in Three Mammalian Genomes. Genome Res. 14, 539–548 (2004).

42. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803–819 (2005).

43. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478, 476–482 (2011).

44. Ponting, C. P., Nellåker, C. & Meader, S. Rapid Turnover of Functional Sequence in Human and Other Genomes. Annu. Rev. Genomics Hum. Genet. 12, 275–299 (2011).

45. Chiaromonte, F. et al. The Share of Human Genomic DNA under Selection Estimated from Human–Mouse Genomic Alignments. Cold Spring Harb. Symp. Quant. Biol. 68, 245–254 (2003).

46. Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 20, 1335–1343 (2010).

47. Smith, N. G. C., Brandström, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. Genomics 84, 806–813 (2004).

48. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? Genome Res. 21, 1769–1776 (2011).

49. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. PLOS Genet. 10, e1004525 (2014).

50. Lunter, G., Ponting, C. P. & Hein, J. Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. PLOS Comput. Biol. 2, e5 (2006).

51. Kellis, M. et al. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. 111, 6131–6138 (2014).

52. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. Genome Res. 17, 1245–1253 (2007).

53. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43, 1031–1034 (2011).

54. Harrow, J. et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012).

55. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. (Wiley-Interscience, 1991).

56. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12,** 41–51 (2011).

57. Kondrashov, A. S. & Crow, J. F. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2,** 229–234 (1993).

CHAPTER 2

Integrating human functional genomic properties using selective pressure

(Gulko B, Siepel A. 2017. Manuscript in preparation.)

2.1 INTRODUCTION

The biology of human genomics can be viewed from two complimentary perspectives: natural selection, which influences the distribution of DNA primary sequence over generations; and functional genomics, which characterizes the immediate biochemistry of DNA within individual tissues. Genomic positions associated with disease and development demonstrate observable patterns of both selective pressure[7–15] and functional properties[1,4–6,59]. However, most sites under selection fall outside protein-coding genes in poorly understood regions of the genome, while most nucleotides evincing biochemical markers (e.g. 80.4% of the genome reported by ENCODE[33]) show no evidence of selective pressure. Over the last ten years, inexpensive sequencing has enabled the collection of vast data sets describing both the distribution of variation across the human species, as well as biochemical properties specific to a variety of tissues[59–61]. While these data sets provide complementary insights that are useful for understanding how genomic properties manifest as observable phenotypes, current methods for integrating them are limited generally sacrificing resolution[62], tissue specificity, or intelligibility[35,63]. This limitation restricts the ability of investigators to integrate expansive contemporary data sets to advance understanding of human molecular genomics, cellular development and personalized medicine[16–19].

Biologists have a pressing need for methods that characterize and rank all positions in the human genome in a manner that provides precision, tissue specificity

and intelligible characterization. Current work in this area can be divided roughly into three categories: (1) unsupervised methods such as ChromHMM[62], Segway[32] and Eigen[63] that seek to optimally encode collections of genomic properties; (2) classifiers such as GWAVA[30] and FunSeq2[64] that attempt to separate highly curated sets of disease associated variants from benign ones; (3) scores representing selective constraint such as phyloP[15], CADD[35], LINSIGHT[65] and fitCons[66] that assess each genomic position's impact on organismal fitness. Methods in the first category can provide informative labeling of genomic positions and suggestions as to which positions may be covered by better-understood genomic properties, but these methods are biased toward explaining variance in the data provided to the algorithm rather than using that data to predict biological implications. Classifiers in the second category identify positions associated with a curated pathology or collection of phenotypes of interest, however these methods do not focus on identifying the diversity of features important for organismal fitness. Such methods provide little information about why one position is given a higher score than another, obscuring the relative importance of differing genomic properties. The third type of method does seek to identify positions influential in organismal fitness, but only LINSIGHT and fitCons characterize genome wide selective pressure, and only FitCons provides sensitivity to genomic properties unique to a cell-type. Methods without this cell-type sensitivity are of limited usefulness in identifying normal patterns of developmental regulation, and by contrast, in characterizing disorders associated with specific developmental stages or specific organs.

Selective pressure at individual genomic positions underlies some of the most successful methodologies for quantifying the importance of individual positions in phenotypic variation. Methods that identify regulatory loci typically use enrichment of conservation scores in identified loci as a methodological validation. However, within

a species selective constraint at individual genomic positions lacks statistical power as it is measured by a depletion in expected variation across individuals while variation itself is typically rare[67,68]. Statistical power in estimating depletion is generally improved by some form of data aggregation. LINSIGHT, for example, aggregates data from over a thousand sources into 49 genomic properties at each position, then further restricts model complexity via a linear combination of those properties used to calculate just two INSIGHT[34] parameters that identify constraint at a position. Alternatively, fitCons identifies cell-type specific activity by aggregating collections of positions according to patterns of functional properties (*functional signatures*) into a collection of related positions called a *functional class*. Selective constraint is then estimated over the collection of positions in each functional class. This particular form of aggregation also shows that individual functional classes can have relatively consistent levels of selective constraint across a variety of cell-types, despite being composed of differing positions in each. However, one limitation of this method is that the native fitCons covariates have relatively low resolution along the genome, generally at the level of 10's to 100's of base pairs (*bp*). In addition, all combinations of covariates are considered, impeding the interpretation of the 100's of functional classes identified. The exponential growth in the number of classes with added covariates also limits the ability of fitCons to include properties with greater precision and variety in types of genomic function.

In the present work, we introduce FitCons2, an inference method based on the same rigorous model of selective pressure as fitCons, but one that scales well with increasing number of covariates and sample cell-types. FitCons2 improves upon fitCons by enhancing genomic resolution, diversity in types of genomic activity, and breath of cell-type specific regulation, while simultaneously simplifying the model to 61 functional classes representing both magnitude and category of genomic activity at

each genomic position in a variety of human tissue types. The central methodology involves development of an intelligible decision tree by recursively identifying and bi-partitioning the individual covariate that is most informative about sites under selective pressure, conditioned on previous partitions[69]. This search-and-partition process is repeated until a statistical significance threshold is reached resulting in a decision tree with a single functional class at each leaf. FitCons2 allows for arbitrary nonlinear relationships among covariates and produces a generative model that simultaneously scores selective pressure and classifies regulatory function for every autosomal position in the human genome. We demonstrate that FitCons2 scores utilized as a classifier are competitive with state-of-the-art prediction methods for identifying sites of disease associated variation according to the National Center for Biotechnology Information (NCBI) ClinVar[50] database and the Human Gene Mutation Database (HGMD)[70]. FitCons2 scores measure differential activity in FANTOM5[71] enhancers across cell-types, and identify active promoter behavior. Furthermore, the cell-type specific scoring reveals regulatory organization derived from tissue type and developmental stage from on a clustering of 115 samples from the NIH Roadmap Epigenomics Project[59]. Finally, the direct interpretation of FitCons2 scores as probability of being under selective constraint provides an estimate of the fraction of the human genome under selective pressure, while also estimating the fraction of selected sites under weak and adaptive forms of selective pressure. Scores, segmentations, and covariates for each Roadmap cell type as well as a cross-tissue aggregate scoring for the hg19 (GRCh37) assembly can be viewed or downloaded via the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser[72].

2.2 RESULTS

2.2.1 Challenges in identifying molecular phenotypes predictive of genomic function

Molecular biologists have devised a range of biochemical assays to identify genomic positions involved in the regulation of gene transcription. A common example is the association of active enhancers with genomic positions bound to histones possessing modifications H3K4me1 and H3K27ac while lacking an H3K4me3 modification[73–75]. Such patterns of assay results are referred to as *molecular phenotypes*. When a molecular phenotype serves as a useful indicator of genomic function we describe it as a *functional signature* and use a suitable collection of functional signatures as an indicator for a property of interest referred to as a *functional class*. With the rapidly growing availability of large collections of genomic measurements, the discovery of functional classes can plausibly be automated by the selection and iterative refinement of informative features. However, the diversity of potentially interesting genomic processes, and uncertainties as to their relative importance, make the automation of functional class identification problematic. Fortunately, evolutionary genomics provides a unifying measure of importance through the mechanism of selective pressure.

A genomic variation in an offspring that changes offspring characteristics can influence its procreative fitness, inducing an observed change in the frequency of that inherited variant over generations. A genomic variation that does not influence procreative fitness is called *neutral* and tends to drift to a differing frequency. The difference between comparable distributions of induced versus drifting variation can be quantified as *selective constraint*. As genomic variation influences selection-sensitive phenotypes **via** transient modifications to biochemical properties of the genome, it is reasonable to consider the amount of selective constraint as a generalized

measure of the importance of coincident molecular phenotypes. Indeed, enrichment for sites under selection is often used as a validation for functional assays[31,33,52,76–78].

There are, however, two major challenges in using selective constraint to identify relevant collections of functional signatures. First, selective constraint operates at a population level over generations, while functional properties change profoundly with tissue-type and developmental phase within an individual. All cells in an organism may contain identical nuclear genetic material, however, genomic activity in adult brain cells, for example, differs widely from genomic activity in adult liver cells and both differ from activity in embryonic cells. Second, selective constraint is difficult to measure. Variation along the human genome is rare (about 1 in 1,000 positions[79]), and selective constraint is generally measured as a depletion in this already rare property, resulting in low statistical power. Variation may also be measured between species as the accumulated impact of millions of years of selective pressure, however such measurements of selective constraint depend on genomic elements maintaining their functional roles over evolutionary time spans. Changes in function at orthologous genomic loci (called *turnover*) may occur with sufficient frequency to degrade interspecies measurements (see section 1.1 Introduction).

2.2.2 Joint inference of functional classes and selective pressure

To address these issues, we introduce FitCons2, a method that identifies parsimonious collections of genomic properties that are maximally informative about expected selective pressure across a broad variety of tissue types. This method is composed of two components: a module that estimates selective pressure for a collection of genomic positions, and a learning model that uses functional signatures to partition the genome into collections of positions that optimize information about distributions of selective pressure. Each of these components is summarized below.

Within FitCons2, selective pressure is modeled by a recently developed statistical method called *INSIGHT* (Inference of Natural Selection from Interspersed Genomically Coherent Elements[34]). The INSIGHT model accepts a collection of dispersed genomic positions and contrasts patterns of polymorphism and divergence at those positions with patterns from nearby neutrally evolving sites, then applies a maximum likelihood inference to estimate the fraction of collected positions under selective pressure (ρ), including those under weak negative (pw) and adaptive (da) selective pressures. The aggregation of data from a large number of sites helps addresses statistical power, while the use of recent primate divergence (chimpanzee, orangutan, and rhesus macaque) as well as polymorphism from 54 unrelated humans[34,80] admits sensitivity to the effects of turnover. INSIGHT also produces a likelihood estimate for the observed data. The associated negative logarithm of the likelihood (*NLL*) for a collection of genomic positions divided by number of positions in that collection, provides an entropy that is used to measure the coherence in selective pressure across the input positions (for details, see section 1.3).

To characterize functional signatures, we identify a collection of nine genomic properties that span a broad range of biological processes and serve as a covariate basis for inference of functional classes (see section 2.4.2 Functional genomic data). Selected covariates cover a range of resolutions, and cell-type specificity including: RNA-seq, DNase-seq, chromatin state, transcription factor binding, and splice site proximity. Cell-type specific covariate values were obtained for 115 karyotype normal cell types from the United States National Institutes of Health (*NIH*) under the Epigenomic Roadmap Project[59]. To identify maximally informative patterns of covariates, the entire genome is recursively partitioned, with each split based on the single covariate that maximizes conditional (post-partition) likelihood under INSIGHT. Each split bipartitions a collection of functional signatures and

consequently, the corresponding genomic positions. The partitioning process repeats on each subset to generate a binary tree, and terminates when the most informative split fails to generate a minimum change in information. For more details on this process see **Figure 2.1**.

The partitioning process results in a decision tree containing 61 leaves (Figure 2.2), with each leaf representing a functional class that in turn corresponds to a collection of molecular phenotypes. Each functional class has a unique INSIGHT model estimating the total, weak, and adaptive selective pressure associated with that class. In a particular cell-type, the functional signature at a genomic position maps that position to exactly one functional class and its corresponding INSIGHT model. The INSIGHT parameter ρ estimates the fraction of positions in a functional class that are under selective pressure. Correspondingly, a larger value of ρ is interpreted as a greater propensity for genomic function, hence, the value of ρ is used as the FitCons2 score for each functional class. All cell-types share the same collection of 61 functional classes, and their corresponding 61 FitCons2 scores. However, a single genomic position can be assigned to different functional classes in different cell-types according to variations cell-type specific molecular phenotype at that position. As a post processing step, scores at each genomic position are integrated across cell-types to generate a single aggregate value for each genomic position (Section 2.2.7 Combining scores across cell types).

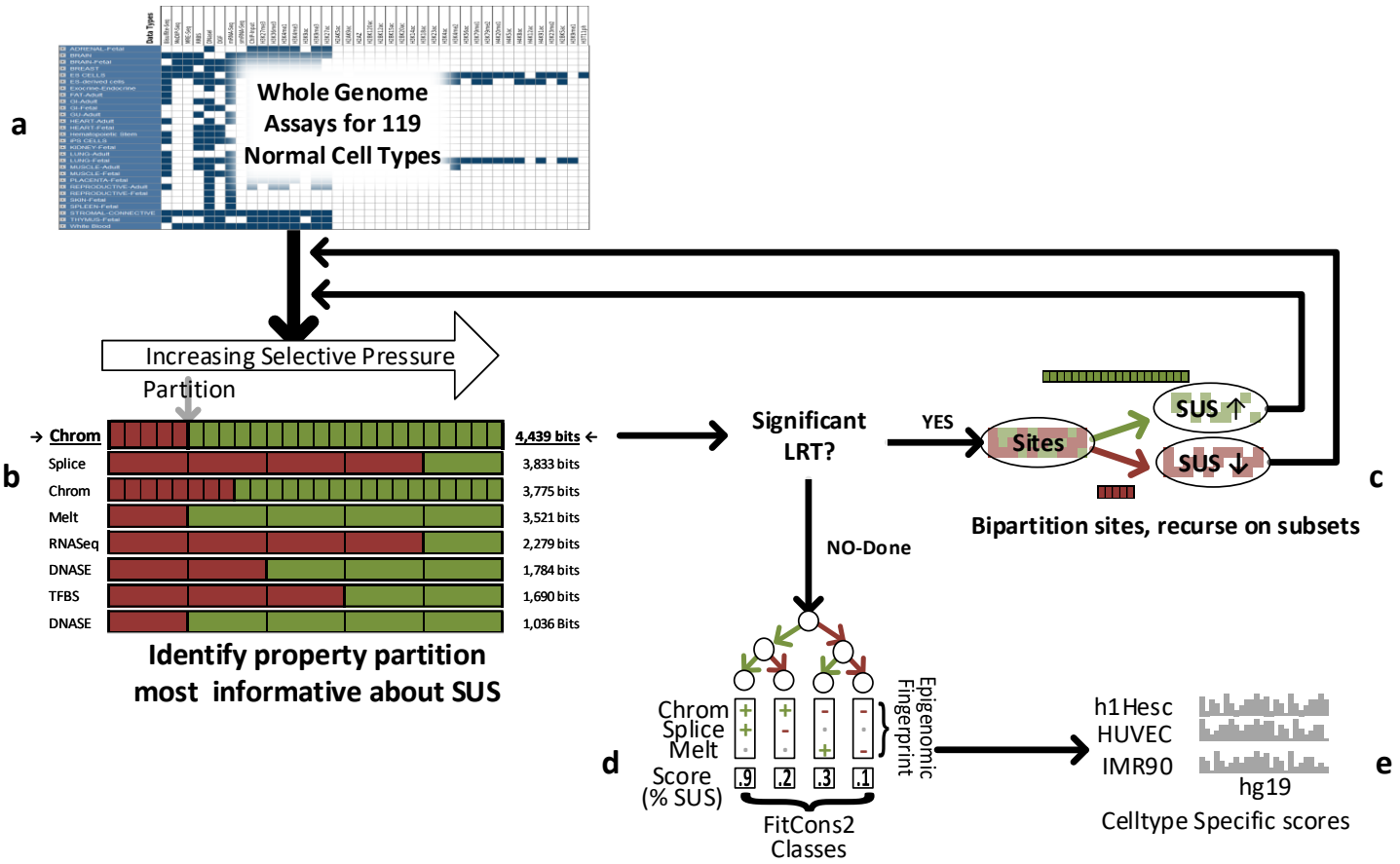
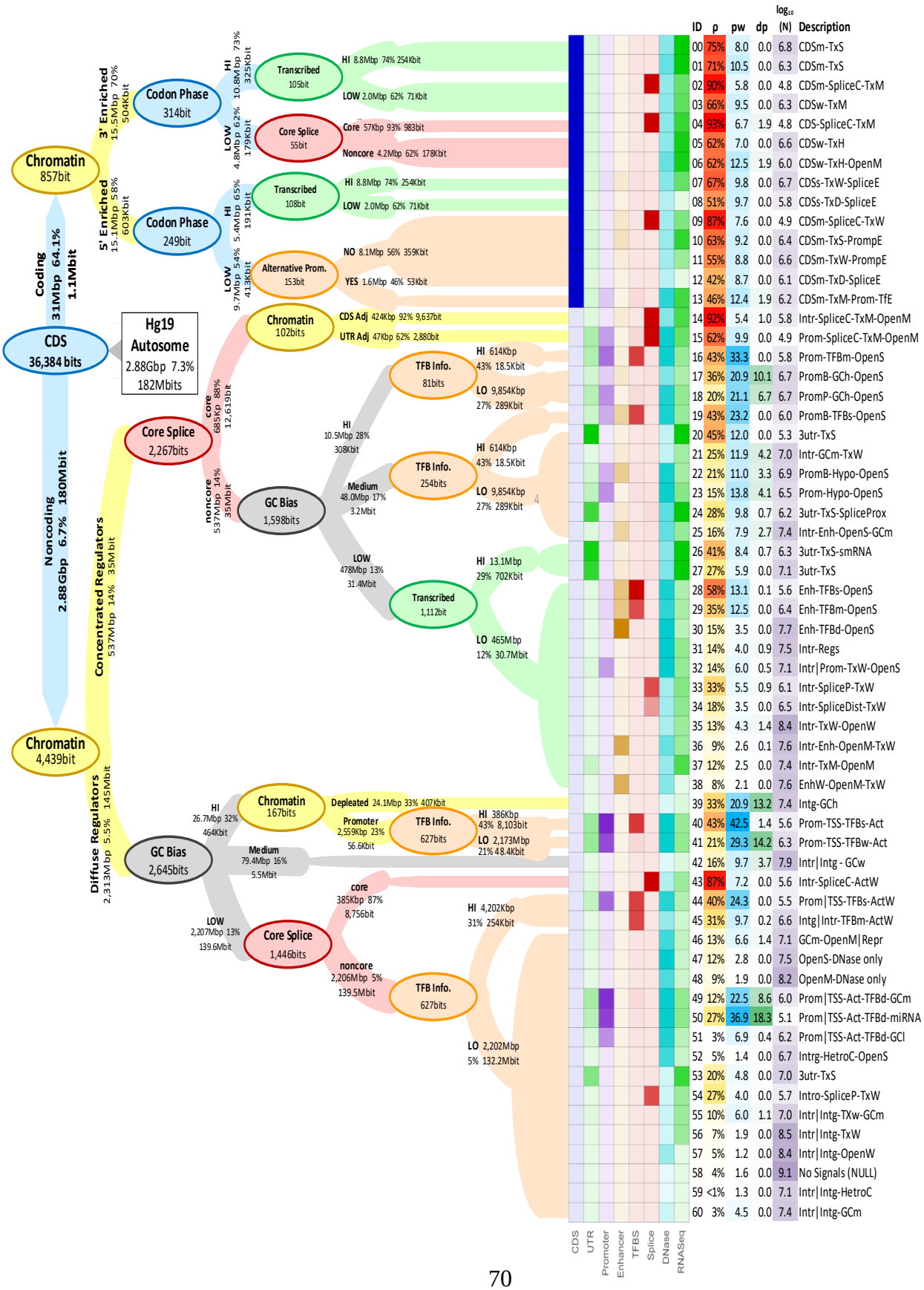


Figure 2.1: Decomposition of covariates into functional classes. (a) Each continuous covariate is quantized into discrete values. (b) At each iteration, non-monotonic values are ordered by increasing selective pressure. Each ordered bipartition of a single covariate is investigated. The partition most informative about likelihood under INSIGHT represents the covariate at the current step. The most informative covariate is selected for bipartition. (c) If the selected split passes a likelihood ratio test (*LRT*), the process is repeated on each product of the bipartition. (d) Once significant splits are exhausted, children of final splits are leaves of a decision tree. Every combination of covariate values belongs to exactly one leaf, with each leaf representing one FitCons2 functional class. A functional class corresponds to one set of INSIGHT parameters. (e) The value of INSIGHT-ρ is the expected fraction of positions under selection in that class, and used as the FitCons2 score for all positions in that class.

Figure 2.2: FitCons2 identifies functional genomic segmentation as a decision tree. The FitCons2 decision tree collects individual genomic properties into functional classes represented by leaves. The hg19 autosome (2.88 billion positions, at root) is recursively split based on the genomic property that is maximally informative about hominin selective pressure under the INSIGHT model. Each internal node is a binary partition of corresponding genomic positions, displaying the partitioning property and improvement in INSIGHT negative log likelihood (*NLL*, in bits) obtained by conditioning on the split. Branches from an internal node describe the corresponding sub-partition's number of genomic positions, fraction of sites under selection, and INSIGHT NLL, averaged across cell types. Recursion terminates when no subdivision results in a statistically significant improvement in likelihood under INSIGHT. Some internal nodes are not represented, to improve visibility. The resultant 61 leaves characterize each genomic position in each of the Roadmap cell types, based the collection of genomic properties that define a functional class unique to that leaf. The relative density of common genomic elements (CDS, UTR, Promoter and Enhancers) is represented in a heat map entry for each leaf, with more saturated color representing greater density. This is followed by a similar display of genomic properties used as FitCons2 covariates: transcription factor binding sites (*TFBS*), Splicing, DNase-seq, and RNA-seq. The final 6 columns provide an ID, expected fraction of sites under selection, as well as expected weak and adaptive substitutions per 10,000 bases, and number of positions in the class followed by a mnemonic name describing the class.



ID	ρ	pw	dp	$\log_{10}$ (N)	Description
00	75%	8.0	0.0	6.8	CDSm-TxS
01	71%	10.5	0.0	6.3	CDSm-TxS
02	90%	5.8	0.0	4.8	CDSm-SpliceC-TxM
03	66%	9.5	0.0	6.3	CDSw-TxM
04	93%	6.7	1.9	4.8	CDS-SpliceC-TxM
05	62%	7.0	0.0	6.6	CDSw-TxH
06	62%	12.5	1.9	6.0	CDSw-TxH-OpenM
07	67%	9.8	0.0	6.7	CDSs-TxW-SpliceE
08	51%	9.7	0.0	5.8	CDSs-TxD-SpliceE
09	87%	7.6	0.0	4.9	CDSm-SpliceC-TxW
10	63%	9.2	0.0	6.4	CDSm-TxS-PrompE
11	55%	8.8	0.0	6.6	CDSm-TxW-PrompE
12	42%	8.7	0.0	6.1	CDSm-TxD-SpliceE
13	46%	12.4	1.9	6.2	CDSm-TxM-Prom-TfE
14	92%	5.4	1.0	5.8	Intr-SpliceC-TxM-OpenM
15	62%	9.9	0.0	4.9	Prom-SpliceC-TxM-OpenM
16	43%	33.3	0.0	5.8	Prom-TFBm-OpenS
17	36%	20.9	10.1	6.7	PromB-GCh-OpenS
18	20%	21.1	6.7	6.7	PromP-GCh-OpenS
19	43%	23.2	0.0	6.0	PromB-TFBs-OpenS
20	45%	12.0	0.0	5.3	3utr-TxS
21	25%	11.9	4.2	7.0	Intr-GCm-TxW
22	21%	11.0	3.3	6.9	PromB-Hypo-OpenS
23	15%	13.8	4.1	6.5	Prom-Hypo-OpenS
24	28%	9.8	0.7	6.2	3utr-TxS-SpliceProx
25	16%	7.9	2.7	7.4	Intr-Enh-OpenS-GCm
26	41%	8.4	0.7	6.3	3utr-TxS-smRNA
27	27%	5.9	0.0	7.1	3utr-TxS
28	58%	13.1	0.1	5.6	Enh-TFBs-OpenS
29	35%	12.5	0.0	6.4	Enh-TFBm-OpenS
30	15%	3.5	0.0	7.7	Enh-TFBd-OpenS
31	14%	4.0	0.9	7.5	Intr-Regs
32	14%	6.0	0.5	7.1	Intr|Prom-TxW-OpenS
33	33%	5.5	0.9	6.1	Intr-SpliceP-TxW
34	18%	3.5	0.0	6.5	Intr-SpliceDist-TxW
35	13%	4.3	1.4	8.4	Intr-TxW-OpenW
36	9%	2.6	0.1	7.6	Intr-Enh-OpenM-TxW
37	12%	2.5	0.0	7.4	Intr-TxM-OpenM
38	8%	2.1	0.0	7.6	EnhW-OpenM-TxW
39	33%	20.9	13.2	7.4	Intg-GCh
40	43%	42.5	1.4	5.6	Prom-TSS-TFBs-Act
41	21%	29.3	14.2	6.3	Prom-TSS-TFBw-Act
42	16%	9.7	3.7	7.9	Intr|Intg - GCw
43	87%	7.2	0.0	5.6	Intr-SpliceC-ActW
44	40%	24.3	0.0	5.5	Prom|TSS-TFBs-ActW
45	31%	9.7	0.2	6.6	Intg|Intr-TFBm-ActW
46	13%	6.6	1.4	7.1	GCm-OpenM|Repr
47	12%	2.8	0.0	7.5	OpenS-DNase only
48	9%	1.9	0.0	8.2	OpenM-DNase only
49	12%	22.5	8.6	6.0	Prom|TSS-Act-TFBd-GCm
50	27%	36.9	18.3	5.1	Prom|TSS-Act-TFBd-miRNA
51	3%	6.9	0.4	6.2	Prom|TSS-Act-TFBd-GCl
52	5%	1.4	0.0	6.7	Intrg-HetroC-OpenS
53	20%	4.8	0.0	7.0	3utr-TxS
54	27%	4.0	0.0	5.7	Intro-SpliceP-TxW
55	10%	6.0	1.1	7.0	Intr|Intg-TXw-GCm
56	7%	1.9	0.0	8.5	Intr|Intg-TxW
57	5%	1.2	0.0	8.4	Intr|Intg-OpenW
58	4%	1.6	0.0	9.1	No Signals (NULL)
59	<1%	1.3	0.0	7.1	Intr|Intg-HetroC
60	3%	4.5	0.0	7.4	Intr|Intg-GCm

2.2.3 Genomic distribution of FitCons2 scores

To obtain a broad overview of the genome-wide distribution of FitCons2 scores, we considered the distribution of scores across the 2,881,033,286 autosomal positions in hg19 for 115 karyotype normal cell-types. As FitCons2 scores are directly interpretable as a measure of selective pressure, the genome-wide average score of 8.19% across positions provides an estimate of the fraction of genomic positions under selection. This breaks down to 19,642,535 selected positions in protein coding regions called *CDS* (64.03%) and 215,714,597 noncoding positions (*NCD*, 7.57%). As has been observed previously, CDS represent a profound concentration of sites under selective pressure, however as GENCODE V19[55] identifies only 1.06-1.18% of the autosome as CDS, a majority of sites estimated to be under selective pressure must be in non-coding regions. FitCons2 identifies 91.7% of the sites that are under selective pressure as being in noncoding regions. This is consistent with the previous studies indicating that a substantial amount of disease associated genomic variation is outside the protein coding genome[81,82].

FitCons2 classes that are highly enriched for specific covariates follow predictable patterns of selective pressure associated with those covariates. The highest scoring positions are found at protein coding splice junctions (CDS class 04, score=0.934 and NCD class 14 Score=0.917), generally within 2 base pairs of the annotated splicing junction. The next highest scoring classes describe CDS (30,678,536 bp) and include classes that span a range of scores from 12:0.423 (class:score) to 04:0.934 with a median score of 0.662 and >90% having a score $\geq$0.514. The highest noncoding scores are associated with classes including highly informative motif positions in transcription factor binding sites (*TFBS*) (28:0.58), then highly transcribed positions (20:0.45), followed by more diffuse promoters (17:0.36), and enhancers (30:0.15). Positions characterized by strong small RNA sequencing

data characteristic of microRNAs are evident in class 50 with a score of 0.27 and the NULL class 58 with no covariate signals has a low score of 0.04.

We examine FitCons2 scores averaged across various previously defined classes of genomic function and also find these to be generally consistent with reported measures of selective pressure[12,40–46]. The core splice sites in introns have an average score of 0.878, while those in exons score 0.787. Protein coding regions (CDS) have an expected score of 0.640. Within a CDS individual units of proteins called amino acids are encode by a series of three based, called a codon. A changed nucleotide in one position can sometimes produce the same amino acid as the original (a *synonymous* substitution). The frequency of this synonymous substitution is correlated with the three positions in a codon, with 4.2%, 1.0% and 66.7% of variants being synonymous in positions 1, 2 and 3 respectively. FitCons2 identifies selective pressure of 0.636, 0.688, 0.594 for codon positions 1, 2 and 3 with increasing selective pressure at codon positions with fewer synonymous substitutions Among noncoding positions, 3'*UTR* (exonic untranslated region) scores have a slightly higher expectation at 0.189 than 5'UTR at 0.185, however the 3'UTR also have a higher variation in scores. The 1,000 bp promoters upstream of transcription start sites have higher average scores at 0.144 than enhancers at 0.103 while the mean intergenic score is 0.063. These estimates for selective constraint are generally consistent with similar measures from other researchers[12,15,83].

We also examined classes of sites identified as active (inactive for the intergenic class) in particular cell types, along with distributions of scores for these sites according to each the respective cell-type's scoring. Distributions for cells from two differing tissue types are displayed below in Figure 2.3. While the specific genomic positions active in each identified class vary substantially among diverse cell types, the genome-wide distribution of scores in each class remains strikingly similar.

This suggests that a position-insensitive relationship exists between commonly identified classes of genomic activity and the spectrum of selective pressures acting on those classes.

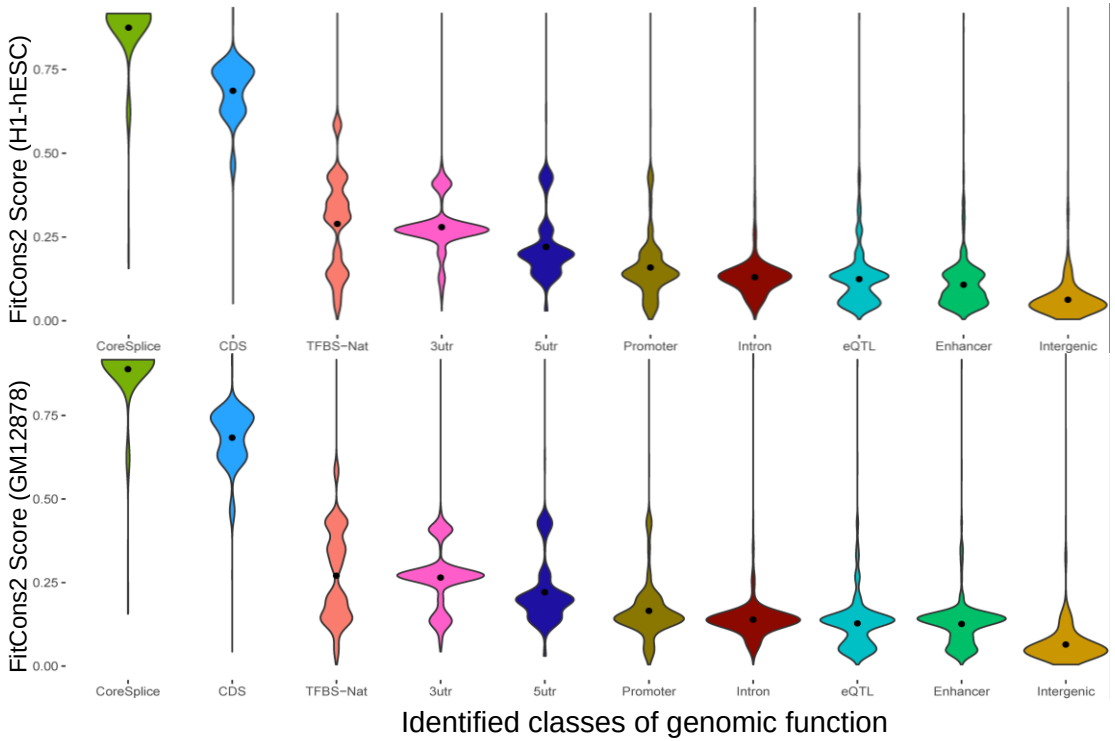


Figure 2.3: Distributions of FitCons2 scores across classes of active functional elements. Distributions of FitCons2 scores in two differing cell types, H1 hESC (*H1*) a human embryonic cell, and GM12878 (*GM*) a B-cell lymphocyte (white blood cell). Each class represents a collection of genomic positions and each position in cell has a score corresponding to its functional signature in that cell-type. Distributions of scores show core splice sites as having a high expected conservation rate of 87.5/88.5% (H1/GM respectively), followed by protein coding sequences (*CDS*) which are 68.6/68.4% conserved, transcription factor binding sites (*TFBS*s), untranslated exonic regions (*UTR*) and 1,000 bp promoters follow respectively. Multimodal distributions such as those visible in CDS and TFBS reveal internal structure like codon phase and positional information in binding motifs. More diffuse classes with intermittent islands of functional activity, such as promoters and enhancers show lower scores than dense classes such as CDS and TFBS.

The shape of the score distribution for a class is evidence of complexities in selective pressure that reflect of biologically significant class substructure. For example, multiple modes in the CDS score distribution reflect effects in the grouping of the three codon positions that identify amino acid encoding degeneracy. Also, scores for transcription factor binding sites (*TFBS*s) divide into modes according to the informative properties of the individual motif positions within a binding site (see Section 2.5.2.10 Transcription factor binding site pseudoannotation). In addition, UTR, promoters and enhancers show peaks of elevated scoring associated with more concentrated interior elements, such as TFBS.

FitCons2 scores provide a quantitative measure for the intuition that more active positions are likely to be under greater selective pressure. Thus, scores for a transcribed CDS class 00 (0.75) are higher than scores for corresponding untranscribed CDS class 11 (0.55). Similarly, scores for DNase I hypersensitive TFBS class 28 (0.58) are higher than scores for the corresponding non-hypersensitive TFBS class 44 (0.40). The higher scores of more active loci suggest the cell type sensitive character the scores, which is visible in the distribution of scores compared across cell types (Figure 2.12,Figure 2.13). Genomic elements such as enhancers follow similar distributions in differing cell types, however, each class represents differing positions within each cell type, based on that cell's biological activity as reflected in covariates such as chromatin state, DNase-seq or RNA-seq. The FitCons2 scores for each Roadmap cell type, their FitCons2 classes, and the covariates that drive them, are available as a track hub viewable in the Cold Spring Harbor Laboratory mirror of the UCSC Genome Browser, and accessible via the FitCons research homepage[84].

2.2.4 Predictive power for regulatory and pathogenic variation

To evaluate the ability of FitCons2 to separate pathogenic from benign noncoding variation, we compared predictive performance for a variety of widely-used

contemporary and classical ranking systems across a variety of frequently referenced classes regulatory and pathogenic variants. Scoring methods included two that were specifically designed to measure selective constraint (phyloP100[85] and GERP++[86]), two that used machine learning approaches to identify features of selection (CADD[35] and LINSIGHT[65]), one targeting pathogenic variations (FunSeq2[64]) and an unsupervised method for organizing a variety of genomic and evolutionary scores (Eigen[63]). The previous generation of this work (FitCons[66]) was also included. To emphasize the genomic resolution of FitCons2, we focused on small functional data sets with loci consisting of a single position (ClinVar & HGMD) or a small collection of positions (transcription factor binding sites, *TFBS*[28]).

FitCons2 scores are designed to identify genomic markers predictive of recent selective pressure. However, genome-wide they also separate sets of annotated functional positions from control sets as well as, or better than, methods specifically designed to predict such features like FunSeq2. To score sites without a clear cell type association, we generated an integrated FitCons2 scoring across cell types. For each position this aggregate combines cell-type specific scores, with a weight representing the information provide by each cell-type, into a weighted cumulative distribution function (CDF) over scores. The FitCons2 decision tree algorithm is run a second time to partition the space of CDFs into 37 classes. Each genomic position belongs to exactly one of these 37 classes while each class, in turn, corresponds to single INSIGHT model with a ρ parameter used as a cell-type integrated (*CTI*) FitCons2 score (section 2.4.7 Cell-type independent score generation).

In TFBS, small loci (>99% are 6-9 bp) are identified by a combination of HUVEC ChIP-seq data and motif matches. FitCons2 scoring of HUVEC provides strong discrimination of these active TFBS from the set of all noncoding sites with nearly 92.4% of these sites having a higher FitCons2-HUVEC score than the lowest

90% of noncoding sites (*NCD*) genome-wide (Figure 2.4, below). The FitCons2 coverage of 92.4% is higher than the 62.6%, 28.4% and 38.1% of TFBS positions identified by FunSeq2, CADD and Eigen at this level of NCD coverage. As illustrated in Figure 2.4.a, the FitCons2 aggregate scoring is nearly as predictive of HUVEC-specific TF binding sites as the cell-type-specific scoring, providing 87.7% coverage of TFBS at 10% coverage of NCD. However, the use of a scoring tailored to the specific tissue type provides even better prediction in the cell type (92.4% TFBS coverage). This effect is also seen in enhancers (section 2.2.6.2 Identification of differentially active enhancers) and suggests the importance of differential scoring to identifying regulators active in distinct phases of cell development (section 2.2.6.1 Tissues cluster by cell type and developmental stage). While the HUVEC ChIP-seq data was part of a data set that contributed to a TFBS annotation used by FitCons2, this annotation was aggregated across many cell types and applied equally to all cell types.

To test the ability of FitCons2 to identify biologically active noncoding positions, we applied a variety of scoring schemes to the HGMD public data set of disease-annotated variants (Figure 2.4.b). FitCons2 covers more than 55.8% of all HGMD single-position variants at a score of ≥0.884 while only 0.034% of NDC positions have this score or higher. Among the observed scores, FitCons2 has the highest coverage of HGMD in the range of 0-5% of noncoding coverage, being then surpassed by LINSIGHT, a related methodology. At only 2.5% noncoding coverage (score of 0.21), FitCons2 identifies more than 71% of HGMD variants (Figure 2.4.b), well above the best performing conservation methods (phyloP, 65% at score 2.011) and comparably to the hi-dimensional LINSIGHT model (69% at score 0.3415).

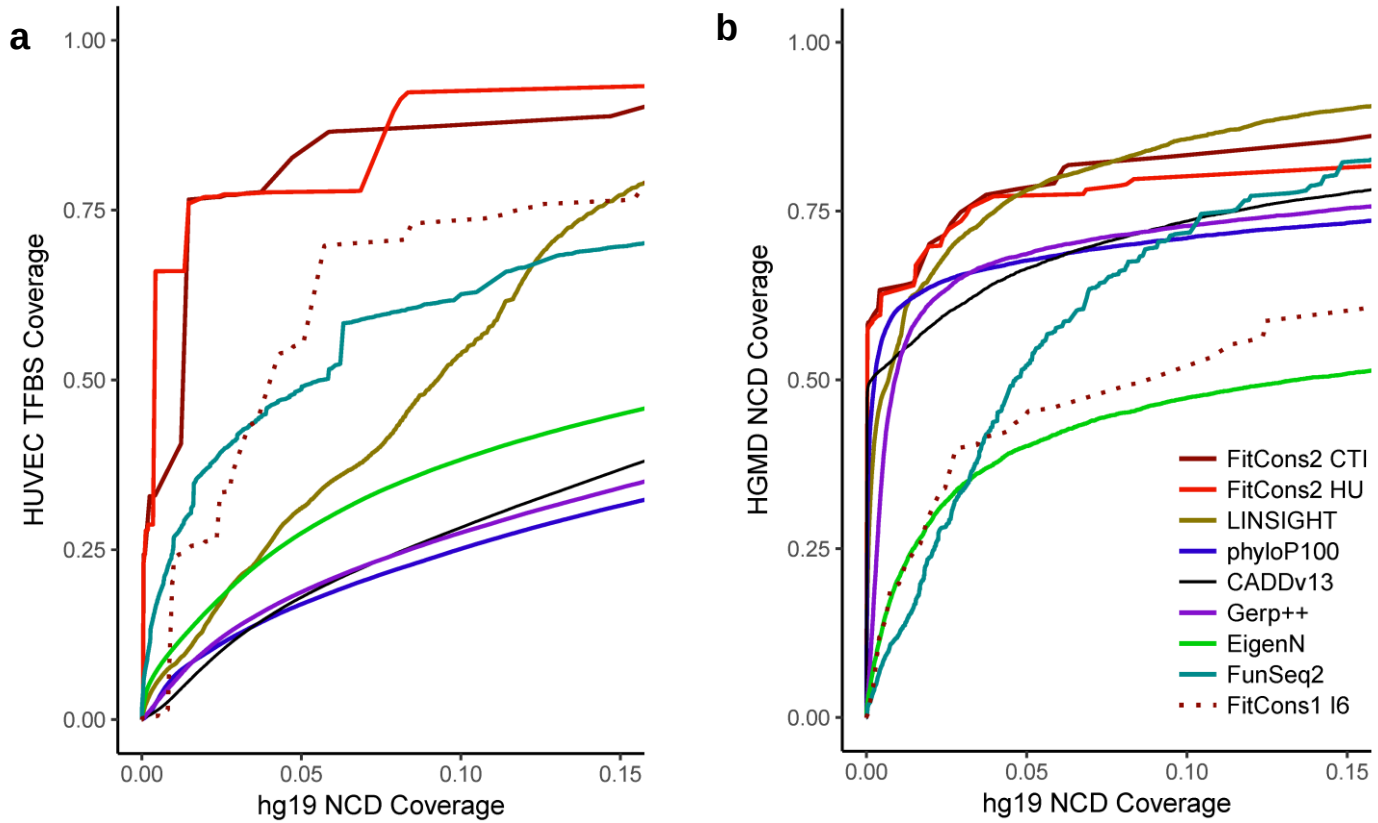


Figure 2.4: Comparative coverage of putative noncoding regulatory elements. FitCons2 is compared to other computational methods utilizing selective pressure to separate transcriptoon factor bindign sites (TFBS) and HGMD pathogenic variants from the 2,846,676,989 genomic noncoding positions (*NCD*). Methods examined include LINSIGHT, phyloP, CADD, Gerp++, Eigen, FunSeq2 and FitCons. Each point on the graph represents fraction of the positive set scoring higher given value (vertical axis) versus fraction of NCD scoring higher than that value (horizonal axis). a) Positive set consists of 57,317 TFBS covering 446,494 positions identified via ChIP-Seq in the HUVEC cell line, versus coverage of all noncoding positions. b) Positive set consists of all 11,591 noncoding autosomal genomic positions annotated as having pathogenic variants in the HGMD database (v89).

We also utilized the ClinVar manually curated database from NCBI which provides both a positive set (pathogenic) and negative set (non-pathogenic) to assess the effectiveness of each scoring system as a discriminator. Despite a reduction in complexity from 624 classes in fitCons to 61 classes in FitCons2, the FitCons2 AUC of 0.962 improves greatly over the fitCons AUC of 0.585. This improvement also demonstrates the central importance of molecular phenotypes informative about diverse genomic functions. The fitCons method had no covariate that was sensitive to the highly conserved intron/exon splice boundaries. Variation at these highly conserved splice sites is a major source of identified noncoding genomic disease in the medical literature. More than 98.8% of the noncoding ClinVar pathogenic variation is in an active splicing class {2,3,4}. Indeed, more than 92.8% of these pathogenic positions are in the highest-scoring core splice class {4}. Alternatively, more than 96.8% of benign positions are in the least active splicing class {1}. Without access to this splicing information fitCons was unable to identify an important biological feature that separates pathogenic variation from benign variation, resulting in a low AUC. However, when splicing features were accessible to FitCons2, they were selected as highly significant despite the restricted number of classes in FitCons2.

The FitCons2 AUC is also well above the 0.772-0.740 generated by Eigen and FunSeq2 and close to the 0.967 to 0.980 demonstrated by methods explicitly using selective pressure as a covariate. For a range of low false positive rate (FPR) values around 2% FitCons2 has the highest accuracy with a true positive rate exceeding 80%. While biases in the selection of ClinVar sites make generalization difficult, scoring breaks down into two broad groups according to AUC, stronger methods with AUC >0.95 and weaker methods with AUC <0.80. The lower group includes Eigen (0.74), FunSeq2 (0.77) and FitCons1 (0.59), while the higher groups includes the remaining methods (0.95-0.98). The higher group includes methods specifically predicting

properties of conservation, including FitCons2. While small differences in AUC have questionable generalizability, FitCons2 is among the top predictors for this data set and identifies the highest fraction of true positives (93%, at a score of 0.57) in a range around an FPR of 2% (Figure 2.5).

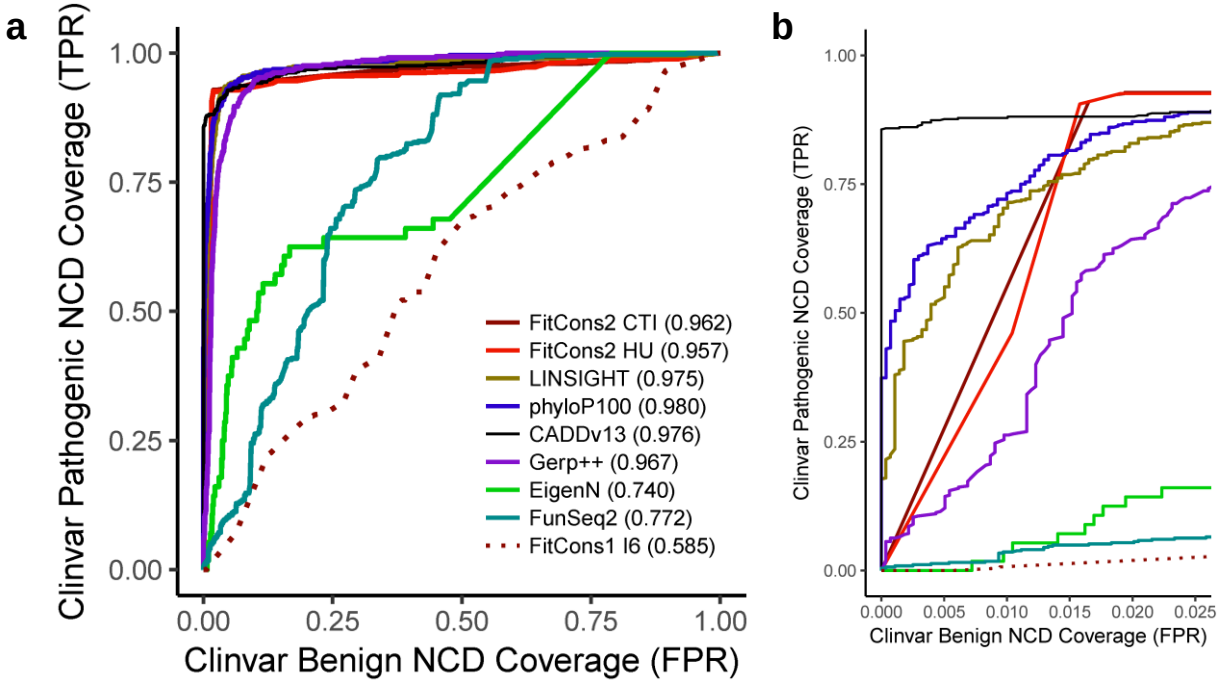


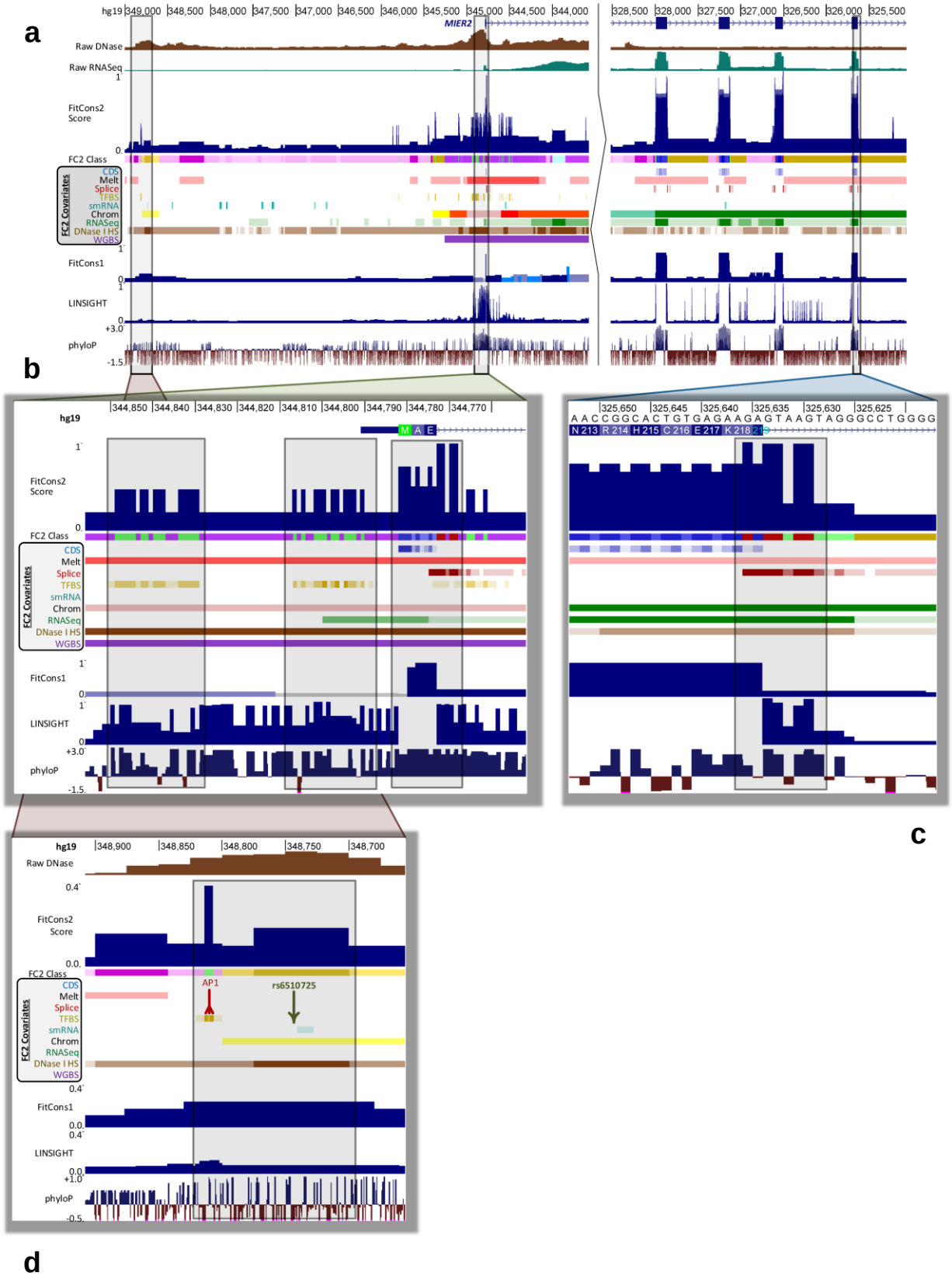
Figure 2.5: Comparison of predictive power on ClinVar variants. FitCons2 is compared to other computational methods utilizing selective pressure to separate annotated non-coding (*NCD*) pathogenic variants from benign ones (control). a) Overview of entire data set with positives being 445 positions annotated as containing pathogenic variants in the ClinVar database, versus 2,785 nonoverlapping positions annotated as containing non-pathogenic variants. This figure takes the form of a Receiver Operating Characteristic (ROC) plot, with each point on a curve representing a classification of each sample in case and control into predicted positives and negatives based on a particular scoring threshold. The data having a score greater than the threshold (vertical axis) is predicted positive with other points predicted negative. The curve is formed by sweeping this threshold across the range of scores and plotting the true positive rate (*TPR*) vs the false positive rate (*FPR*). The Area Under the Curve statistic (AUC, legend) provides a measurement of prediction accuracy, ranging from 0.0-1.0, with a value of 1.0 as perfect classification and 0.5 being random. b) Detail of left 2.5% of range shows relative performance of methods at very low FPR, a common area of interest to researchers. FitCons2 provides the highest TPR of all methods at an FPR of around 2%.

FitCons2 performs approximately as well as competitive methods in identifying functional sites employing a readily interpretable 61 class genomic segmentation and without the use of selective pressure as a covariate. Despite this simplicity, FitCons2 is competitive with or exceeds more complex contemporary models in identifying ClinVar and HGMD positions and is the most powerful predictor of HUVEC TF binding at a range of low false positive rates.

2.2.5 Resolution and interpretation of functional classes

To visualize the interplay among covariates, classes, and scores near the MIER2 gene, these properties were displayed using the UCSC genome browser (Figure 2.6). This display provides a graphic representation of the relationship between a putative upstream enhancer, promoter / transcription start site (*TSS*), and an internal exon. Within a DNase-seq raw signal peak characterized by enhancer associated chromatin marks (Figure 2.6.a at left, and d), FitCons1 scores shows a broad activity peak, while the more focused FitCons2 peaks correspond to LINSIGHT scores. However, in this region FitCons2 also clearly identifies substructures such as a TF binding site for AP1. AP1 is a known regulator for MIER2, and the binding locus is localized by a 7bp spike in scores (FitCons2 class:score of 45:0.31). Downstream of this binding site is a lower intensity peak (30:0.15) identified as associated with an oncogenic variant (rs6510725[87]) next to a small RNA-seq signal (light blue) and DNase-seq peak (brown), suggestive of eRNA from an enhancer. The phyloP score for this region is not elevated, indicating that this may be a recently evolved regulator.

Figure 2.6: Detail of MIER2 gene and upstream loci. A UCSC Genome Browser display of the MIER2 gene in the HUVEC cell line, including (a) overview, (b) details of the transcription start site, (c) an exonic splice site, and (d) a candidate upstream regulatory locus for the MIER2 Gene. Hg19 positions follow the horizontal axis between chr19 325,000-349,000. Rows delineate properties varying across positions while callouts detail relevant subsets of rows with greater horizontal detail. The overview rows (a) begin with a genomic position, followed by MIER2 exons (in blue), two raw covariates DNase-seq (log scale) and RNA-seq signals, the FitCons2 score, FitCons2 class and all quantized covariates. Darker covariate colors generally represent more intense signals indicating a greater impact on FitCons2 score. Following the covariates are the fitCons, LINSIGHT and phyloP100 scores. Also shown in detail d) is a functional variant (red) and an identified transcription factor binding site (green).


a
b
c
d
hg19
MIER2
Raw DNase
Raw RNASeq
FitCons2 Score
FC2 Class
FC2 Covariates
CDS
Melt
Splice
TFBS
smRNA
Chrom
RNASeq
DNase I HS
WGBS
FitCons1
LINSIGHT
phyloP
AP1
rs6510725

As coordinates approach the transcription start site (*TSS*), selective pressures appear to come from more ancient sources as FitCons2, LINSIGHT and phyloP scores become more consistently positive. Immediately upstream of the TSS, FitCons2 identifies two highlighted signal peak clusters associated with the promoter. The internal structures in these score peaks reflect motifs at individual transcription factor binding sites (gold covariate) with increased selective pressure found at higher information content motif positions (16:0.43) and less selective constraint at less informative motif positions (18:0.20). The start of the protein coding region is highlighted by a peak in scores signaling higher levels of conservation at the start codon, visibly labeled as **M**. The subsequent alterations in score suggest differential selective pressure indicative of each codon's amino acid degeneracy substructure. Score variation also highlights a peak at the core exonic splice sites, and another at intronic positions adjacent to the splice junction (07:0.87 and 14:0.92, respectively). The distance to an active splice site represents one of the most complex and powerful indications of selective constraint in FitCons2. A metaplot of INSIGHT-ρ vs distance from splice sites shows wide variations in expected values from 0.93 to 0.57 and back to 0.76 within 5 bp downstream of the 5' splice site (see Sup. Figure 2.18). At the 3' splice site this variation is even more pronounced traversing 0.97 to 0.15 to 0.45 within 5 base pairs of the end of an intron. In HUVEC the highlighted exon shown in sub-panel (d) is transcribed and the intronic splice site just downstream of the exon belongs to class 14 with score of 0.92. However, in cell types like GM12878 this same position has a low RNA-seq signal, and the reduced covariate activity puts this position into class 41 producing a score of only 0.21.

Within a three base-pair codon, some DNA single nucleotide substitutions result in identical amino acids when translated into a protein. Such variations are referred to as synonymous substitutions and generate little phenotypic impact hence

have low exposure to selective pressure. The fraction of possible synonymous variation cycles with codon phase, and the resultant changes in selective pressure are visible as a regular exonic crenulation structure in subpanel (c). Generally, the central codon position is under the greatest selective pressure, followed closely by the first with the third being substantially less, and this tendency is indeed visible in HUVEC scores where the first and second phases group together in class 01 of the FitCons2 decision tree (Figure 2.2) with a score of 0.71 but differentiate from the third which belongs to class 06, with score 0.62. In an adjacent codon, identified as K218, FitCons2 identifies splicing proximity as more informative about selective pressures than codon phase, resulting in an assignment of the third position to splicing-associated class 04 with an elevated score of 0.93 rather than the lower scoring transcribed CDS-associated class 06:0.62 .

As the illustration of MIER2 demonstrates, FitCons2 infers genomic properties predictive of selective pressure, and characterizes positions based on relevant epigenomic activity. Both coding and noncoding scores are represented in terms of the same underlying measure, INSIGHT-ρ. The result is a classification system that segments the genome according to biological activity, but also weighs the relative importance of differing patterns of properties. This classification identifies simple combinations of genomic properties that are the most descriptive of the biological activity exposed to selective pressure at each genomic position.

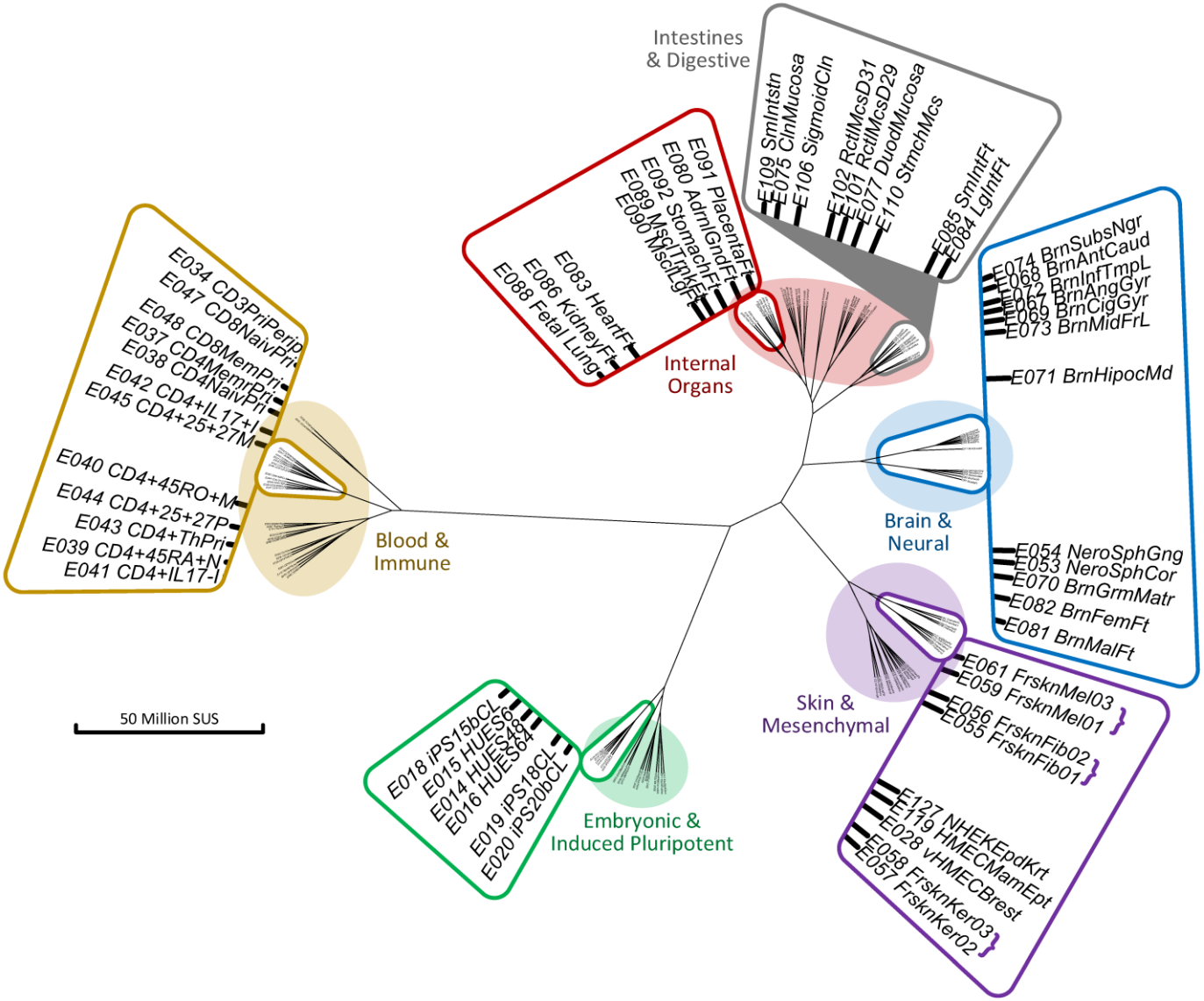
2.2.6 Characterizing tissue-specific genomic activity

2.2.6.1 Tissues cluster by cell type and developmental stage

Quantifying the cell-type specific regulatory activity that influences procreative fitness is of central interest to researchers in human epigenetics. However, quantification efforts have generally been limited to assessments of a small collection of properties such as transcription, chromatin accessibility, or enhancer activity. To

investigate the relationship between activity in differing cell types, we explored a clustering of 115 cell types based on differences in scores across all autosomal genomic positions in cell-type pairs (Figure 2.7). The combination of a broad-based quantification of activity based on selective pressure with cell-type sensitivity provides a unique view into similarities and dissimilarities among cell types. Distance between cell types is measured by summing the absolute value of scores across all positions in the genome to form an L1 or "Manhattan" distance. As the score at each position represents a probability of being under selective pressure, the natural distance metric is in units of expected number of sites under selection (*SUS*) across the hg19 reference autosome. The minimum extent of any cell type in this space was 223,019,006 SUS and the maximum 264,070,590 SUS, with a median of 234,392,692 SUS. The minimum of the 6,555 distances between pairs of differing cell-types was 11,433,006 SUS and the maximum 61,602,121 SUS. The distance matrix was clustered using the default Ward-D2 clustering method on the R package V3.3.1. Projection from a 115-dimensional space to a 2-dimensional tree induces some distortion, making small distances appear smaller and large distances appear larger; however, some important relationships are clearly visible.

Figure 2.7: Tissue and developmental states cluster by FitCons2 scores. Clustering of 115 Roadmap cell types by variation in FitCons2 scores across all genomic positions shows strong tissue specificity. Immune and blood cell types cluster together (gold) and among them T-cells (outset, lower) group particularly tightly. Brain and neural cells (blue) show similar patterns of activity, and in particular fetal brain tissues (blue outset, bottom) cluster separately from adult tissues (blue outset, top). Similarly, fetal organ tissue (red outset) cluster within the broad cluster of internal organ tissues (red). Digestive tissue samples are shown in a grey outset, while the purple cluster contains skin and mesenchymal cell types. Tissue replicates of Fibroblasts, Keratinocytes and Melanocytes are immediately adjacent (purple outset, curly braces). Embryonic cells and induced pluripotent stem cells cluster tightly (green, outset) within the broader grouping of less differentiated progenitor cell types (green). While the projection to 2 dimension distorts distances, this figure has a natural distance scale in units of $\sum_{i \in hg19} |\Delta \text{INSIGHT}\, \rho|_i$, that is, hg19 position-summed differences in expected sites under selection. Broadly, this allows changes in epigenetic properties between differing tissue types within a single organism to be interpreted in a scale of selective pressure concordant with evolutionary turnover.


Intestines
& Digestive
E109 SmlIntstn
E075 ClnMucosa
E106 SigmoidCln
E102 RctlMcsD31
E101 RctlMcsD29
E077 DuodMucosa
E110 StmchMcs
E085 SmlIntFt
E084 LgIntFt
E091 PlacentaFt
E080 AdrnlGndFt
E092 StomachFt
E089 MscLTrnkFt
E090 MscLLgFt
E083 HeartFt
E086 KidneyFt
E088 Fetal Lung
Internal
Organs
E074 BrnSubsNgr
E068 BrnAntCaud
E072 BrnInfTmpL
E067 BrnAngGyr
E069 BrnCigGyr
E073 BrnMidFrL
E071 BrnHipocMd
Brain &
Neural
E054 NeroSphGng
E053 NeroSphCor
E070 BrnGrmMatr
E082 BrnFemFt
E081 BrnMalFt
E034 CD3PriPerip
E047 CD8NaivPri
E048 CD8MemPri
E037 CD4MemPri
E038 CD4NaivPri
E042 CD4+IL17+I
E045 CD4+25+27M
E040 CD4+45RO+M
E044 CD4+25+27P
E043 CD4+ThPri
E039 CD4+45RA+N
E041 CD4+IL17-I
Blood &
Immune
50 Million SUS
E018 iPS15bCL
E015 HUES6
E014 HUES48
E016 HUES64
E019 iPS18CL
E020 iPS20bCL
Embryonic &
Induced Pluripotent
Skin &
Mesenchymal
E061 FrsknMel03
E059 FrsknMel01
E056 FrsknFib02
E055 FrsknFib01
E127 NHEKEpdKrt
E119 HMECMamEpt
E028 vHMECBrest
E058 FrsknKer03
E057 FrsknKer02

Cell types gather into clusters readily associated with tissue types, corresponding to embryonic (green), blood (gold), neural (blue), connective tissue (purple) and internal organs (red & gray). As expected, replicate cell-types such as the pairs of fibroblasts, keratinocytes and melanocytes are immediately adjacent, showing highly similar patterns of FitCons2 scores (purple outset, braces). Strikingly, some cells cluster by differentiation stage while others cluster by related organ. Thus, induced pluripotent stem cells show similar patterns of genome wide activity as embryonic stem cells (green), while brain and neural cells cluster together regardless of developmental stage (blue). However, the induced H9 Neuronal Precursor cell line clusters with ESC's, while within the neural tissue cluster embryonic brain tissues cluster together (blue outset, at bottom) and separately from adult brain tissues (blue outset, at top). The fetal neural tissue in the Brain & Neural cluster was sampled at 17-20 weeks after gestation, suggesting a regulatory phase change between embryonic stem cells and differentiated fetal neural cells before 17 weeks of age. Similarly, fetal organ (red, outset) and fetal digestive tissues (grey outset, right) cluster together within their respective tissue types.

2.2.6.2 Identification of differentially active enhancers

We evaluated the ability of FitCons2 to distinguish between active and inactive states of specific regulatory elements by comparing a shared set of 1,026 FANTOM5 enhancers covering 375,480 genomic positions with differential activity among three cell types, GM12878, HUVEC, and H1 hESC. FANTOM5[71] enhancers are small loci (mean 366 bp, st. dev. 201 bp) identified using Cap Analysis Gene Expression[88] (CAGE) to locate regions with divergent transcription that is associated with enhancers. Differentially active enhancers are defined as those among the top 10% of CAGE read depths in at least one of the three cell types (*active*), and zero read depth in at least one of the three cell types (*inactive*). Enhancers that are neither active nor

inactive in a particular cell type are removed. For each cell type, the mean FitCons2 score for enhancers in each class is calculated using the relevant cell-type specific scoring. The mean score across positions for active elements in a cell type was consistently higher than the mean score across inactive elements (Figure 2.8), demonstrating sensitivity to the more highly active enhancers in each cell type.

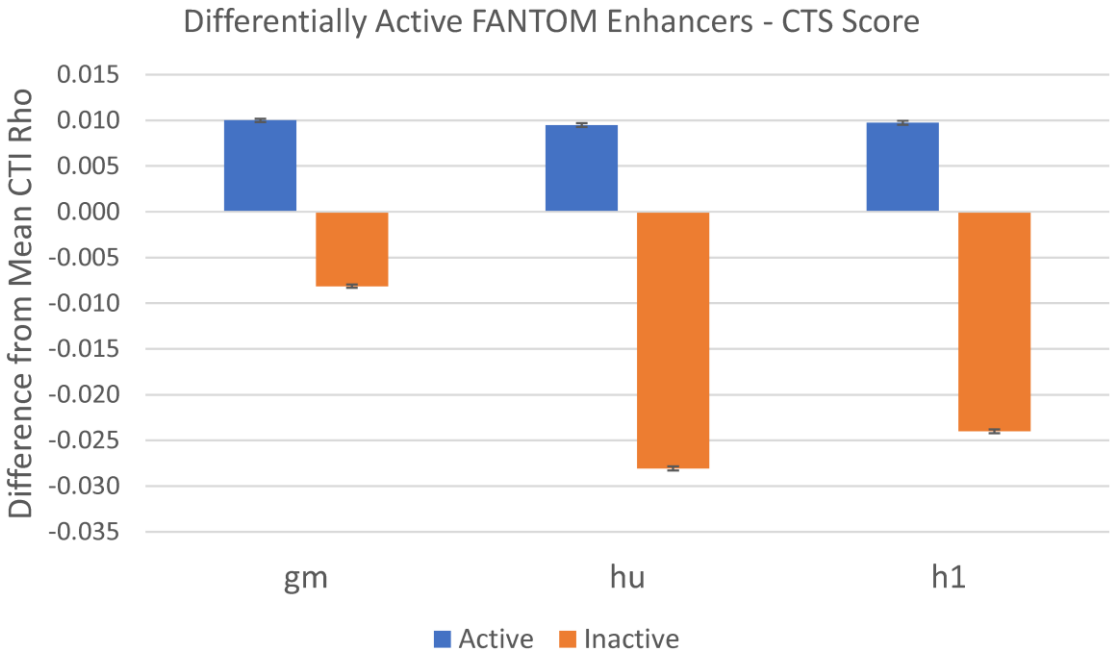


Figure 2.8: Tracking enhancer activity across cell-types. Mean FitCons2 scores of differentially active enhancer across three cell types: GM12878 (*gm*), HUVEC (*hu*) and H1 hESC (*h1*). Mean cell-type specific scores of active enhancers are uniformly higher than mean scores of enhancers inactive in the same cell-type. The same set of enhancers is used in all three cell types, but the activity of each individual enhancer varies by cell-type.

To further characterize the predictive power in cell-type specific FitCons2 scorings, the distribution of scores for each combination of cell-type and activity state was generated and used to produce three ROC plots. These plots comparing the relative discriminative power of the same-cell scoring with off-cell scoring of the cell-type sensitive differential enhancers (Figure 2.9). In each plot, the true positives are positions in an active enhancer, while the true negatives are positions in an inactive

enhancer. Each enhancer is active in at least one of the three cell types, and inactive in at least one of the others. In each case the same-cell score provided a higher AUC, indicating a better predictive accuracy for active enhancers. The mean same-cell AUC was 0.70, while the mean off-cell AUC was 0.46, no better than random assignment.

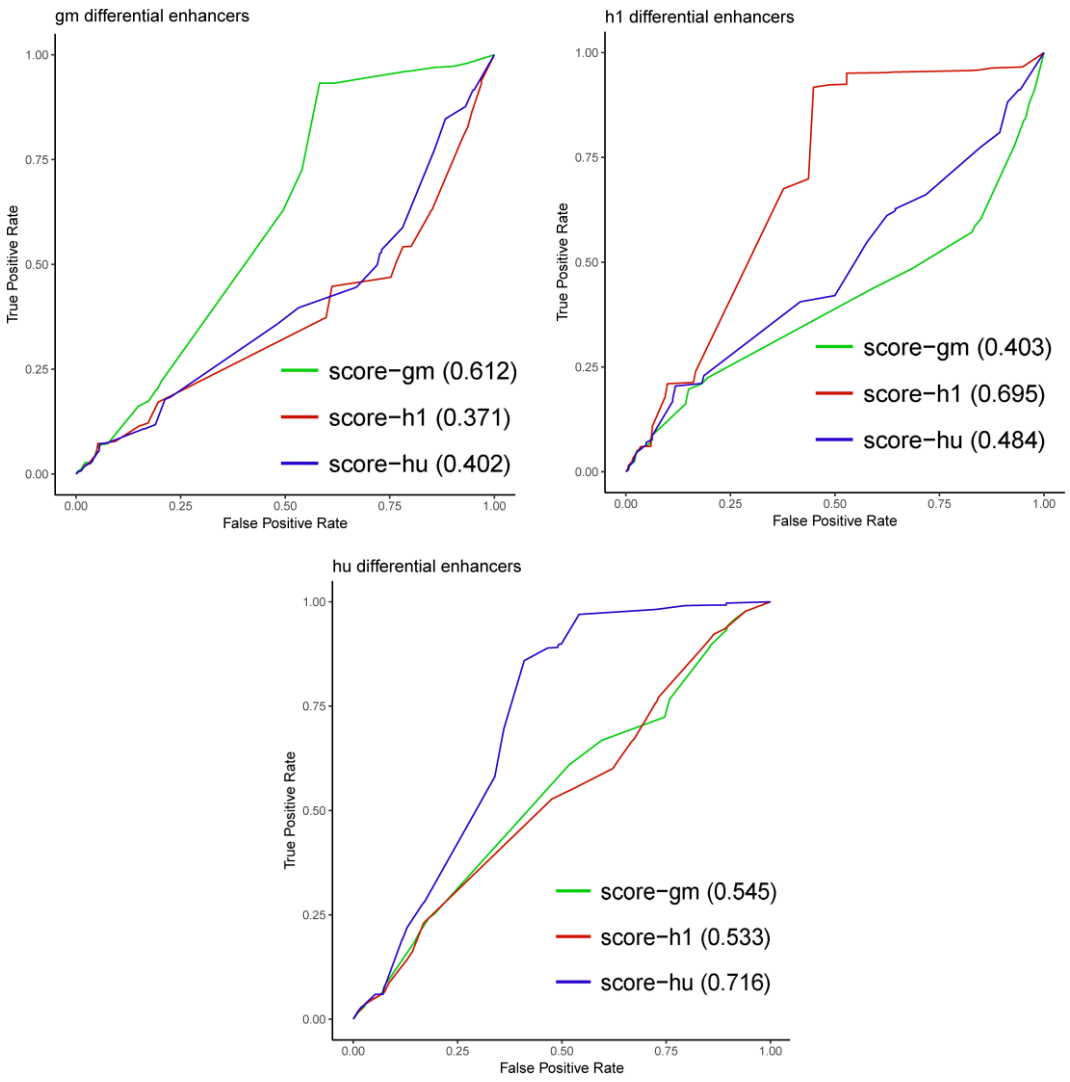


Figure 2.9: ROC plots showing FitCons2 scores tracking enhancer actitivy across cell-types. For a single collection of differentially active enhancers, positions in enhancers active in a particular cell-type are the true positive, while positions in inactive enhancers are the true negatives. Each of the three panels represents FANTOM5 enhancer activity in one cell-type (GM12878, HUVEC, or H1 hESC), but is assessed using scores from each of the three cell types. Scores from cell-types matching the active enhancer cell-type show predictive power, with AUCs of 0.61 to 0.72. Scores from unmatched cell-types show little predictive power with AUCs of 0.37-0.55. The score variation used to discriminate active from inactive enhancers positions occurs at intermediate false positive rates because enhancers are relatively diffuse structures with a variable density of active positions and contain many low scoring positions even in the active state.

2.2.7 Combining scores across cell types

FitCons2 scoring was developed on a karyotype-normal subset of 115 Roadmap cell types, and scores were subsequently generated for all 127 available cell-types representing the 19 Roadmap tissue group classifications. It has been demonstrated that functional properties of tissues cluster by cell type. Consequently, the most representative scoring for a position is likely to be the one trained on a similar Roadmap cell-type. However, it is not always practical for investigators to identify the effective tissue type and developmental stage for a genomic locus of interest. What is needed is a measure that consolidates all cell-types into a single scoring of human genomic positions. Obvious aggregation methods such as max or mean score across cell types degrade either dynamic range (mean) or interpretation as expected probability of selective pressure (max). Furthermore, as Roadmap includes more of some cell-types than others, it was not clear how to weigh combinations of high and low scores in the presence of cell-type selection bias.

To address these issues, a weight (w) for each cell-type (n) was generated that reflected each cell-type's unique variation as well as a proportionate fraction of the variation shared with other cell-types (w_n, see 2.5.6.1 Cell-type information weighting). The problem of integration scores across cell-types at each genomic position i could then be framed as a mapping from a collection of 115 cell-type (n) specific weighted scores ($n \in [1,115]$, $\{w_n, \rho_{n,i}\}$) to a single real number $\boldsymbol{\rho}_i$. As relationships between selective pressure ($\boldsymbol{\rho_i}$) and collections of $\{w_n, \rho_{n,i}\}$ might be complex, this problem resembled the original mapping of complex genomic features into selective pressure. Under the assumption of exchangeability across n among collections of $\{w_n, \rho_{n,i}\}$, we were able to reuse the FitCons2 framework as follows:

- A set of 12 levels ($v \in [1, \ldots, 12]$) of ρ was selected to represent a quantization of all scores across all cell-types (ρ_v). This was done in a

manner similar to the selection of covariate quantization boundaries in the cell-type sensitive scoring (see 2.4.1 Covariate generation).

- A separate covariate c_v was generated for each of the 12 levels. For a genomic position i, the numerical value for a covariate ($c_{v,i}$) was the sum of the weights of all cell types (n) with a score above the corresponding boundary ρ_v, specifically, $c_{v,i} = \sum_{n \in N} w_n \mathbf{1}_{(\rho_{n,i} \geq \rho_v)}$. This step aggregated weights across cell-types.

- Each real-valued covariate was now quantized, as with other real-valued genomic properties (like RNA-seq read depth, see 2.4.1) into 5 levels.

- FitCons2 was run on this collection of 12 quantized covariates, generating a new decision tree.

- Just as with the cell-type sensitive segmentation, each leaf has a score (ρ) generated by INSIGHT as an estimate of the fraction of sites under selection. Similarly, each genomic position belonged to exactly one of the cell-type insensitive classes and is assigned that class's score.

The result was a set of 37 cell-type independent (*CTI*) classes, each with an INSIGHT-ρ values ranging from 0.038 to 0.884. This aggregate scoring was used for analyzing data sets like ClinVar and HGMD, where the most relevant cell type may be impractical to uncover. While conceptually similar to a maximum score across cell-

types at each genomic position, each aggregated score was the result of an INSIGHT calculation thus carried the interpretation as a probability of being under selection. This CTI scoring can be used in an initial screening for regions of likely genomic interest and for comparison to scores like LINSIGHT, phyloP, and Eigen which do not discriminate based on cell-type, see Figure 2.4 and Figure 2.5.

2.2.8 Deconstructing the Coordinator motif

To demonstrate an application of FitCons2 cell-type sensitivity, we employ FitCons2 scores to help characterize binding events at a recently identified 17 base-pair primary sequence motif called Coordinator[89]. The Coordinator motif was discovered in a study of enhancers influential in the epigenetic divergence of human and chimpanzee cranio-facial (*CF*) features. The presence of this motif was found strongly predictive of surrounding chromatin features and is speculated to play a role in the development of enhancer competence. However, the specific transcription factors that are active at this site during craniofacial development are not well understood. Of 78 available transcription factors (*TF*s), 72 bind in some instance of this motif in a CF enhancer. To identify TFs most influential in CF development at Coordinator, we employ cell-type sensitive FitCons2 scores to separate developmental enhancers from constitutive ones. We then highlight TFs preferentially found at Coordinator motifs in those developmental enhancers as candidates for biological validation.

Coordinator[89] was identified as a statistically enriched motif found in a collection of 14,153 short autosomal CF enhancer loci with a mean size of 200 bp. These enhancer elements have been identified as a source of phenotypic variation between modern humans and the most recent common ancestor with chimpanzees. Variations in the 17 base-pair Coordinator motifs found in CF enhancers are reported to confer large phenotypic effects, however, the specific transcription factors driving

biochemical activity are not well understood. To create a baseline, a set of the most significant 13,480 Coordinator motif matches (*hits*) in CF enhancers were calculated, corresponding to approximately one hit per identified enhancer. Next, a reference database[28] of transcription faction (TF) binding positions for 78 known TFs was intersected with these 13,480 hits, referred to as enhancer motifs. The number of identified binding events for each TF was determined at each corresponding position in the Coordinator motif. Prolific binding was observed, with 72 of the 78 TFs found at least once, 63 at least twice, and 13 bound 20 times or more.

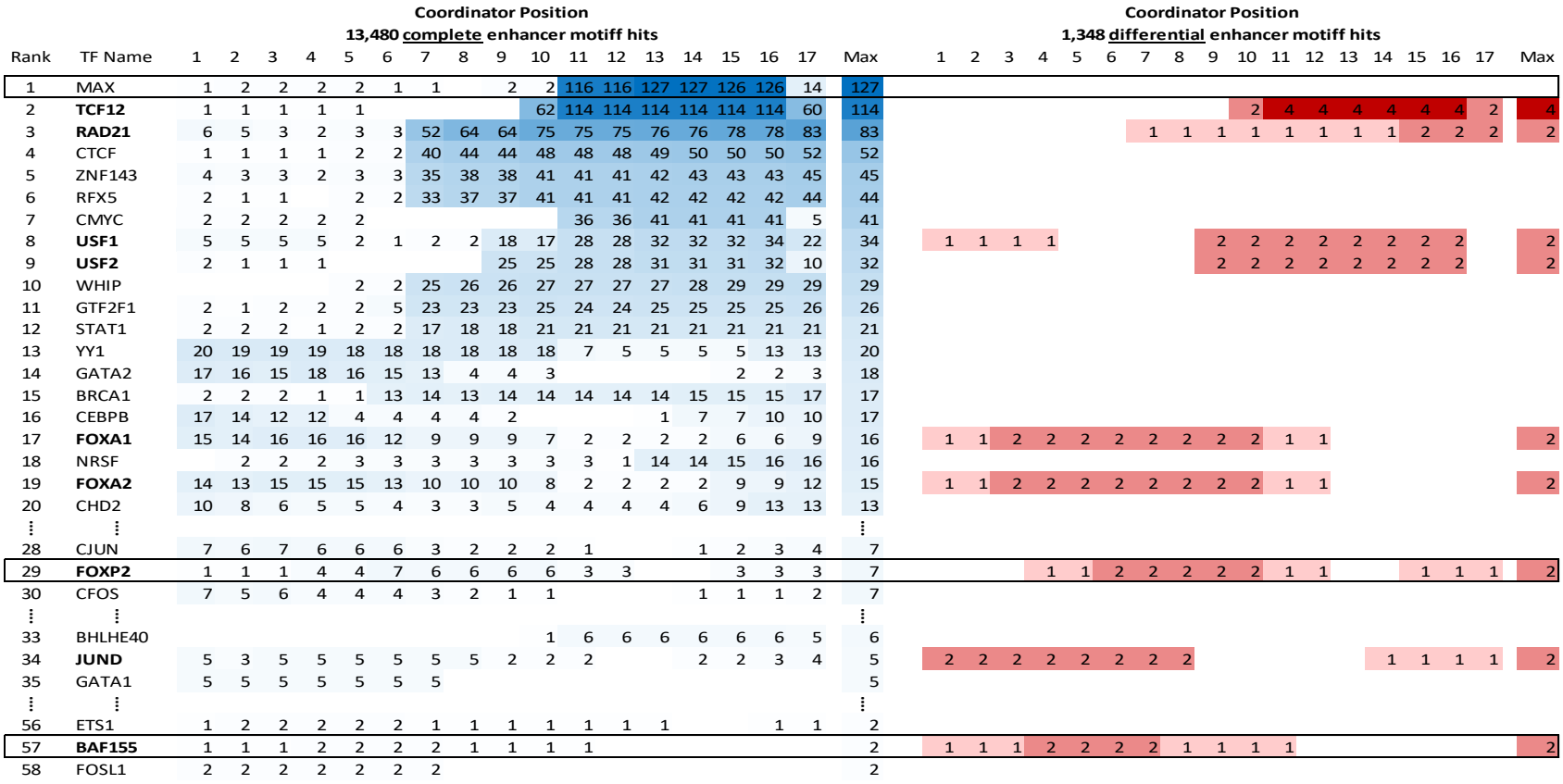
Developmental enhancers are those with activity levels that change with cellular development. To identify developmental TF binding associated with developmental CF enhancer activity, the average FitCons2 score in each of the 13,480 enhancers motifs was calculated in early developmental cell-types (embryonic cells) and more differentiated cell-types (neural and chondrocytic). The enhancer motifs with the greatest score differential between an embryonic and an adult cell type were taken as enriched for developmental activity. The enhancer motif hits with the top 10% (1,348) of score differentials were taken as *developmental enhancer motif*s. These developmental enhancer motifs were then intersected with the TF database to identify the number of TF binding events at each of the 17 positions in the developmental coordinator motifs.

The distribution of TF hits in the developmental enhancer motifs was strikingly different from what might be expected from a random sample of all enhancer motifs (Figure 2.10: Transcription factor binding in enhancer Coordinator motifs), with only 24 TF's having at least 1 hit in this developmental set, 9 with 2 or more hits, and only one with more than two hits. The most abundant TFs on the complete set were relatively depleted, or completely absent, from the developmental set. In particular MAX, the most abundant TF in the complete set with 127 hits, was completely absent

in the developmental subset, while the next three most abundant TFs; TCF12, RAD21 and CTCF were found at 2.9, 4.2, 5.2-fold depletions. The most enriched TFs in the developmental set, FOXP2, JUND, and BAF155, were found at 2.9, 4.0 and 10.0-fold enrichments. In particular, both of the BAF155 hits found in the complete set of enhancer motifs were also found in the developmental set (Figure 2.10, highlighted).

Information about binding in the Coordinator motif is limited. Higher numbered positions are reported resemble an E-box (enhancer box) with HOX-like motifs that facilitate the initiation of transcription. Indeed, known E-box associated TFs USF1/2 and TCF12 are found in higher numbered motif positions (right) in both complete and developmental enhancer motifs. However, binding at the lower numbered motif positions (left) is less understood. The most developmentally enriched TF, BAF155, is enriched on the left side of Coordinator and is associated with neural development via participation in nBAF / npBAF protein complexes. Variations in this protein are also associated with facial deformation and hirsutism in humans[90,91] (UniProt Q92922[92]). FOXP2 is also developmentally enriched at the lower Coordinator positions. FOXP2 is associated with neural development and in particular, brain formation during embryogenesis. Variation in this protein is well known to be associated with orofacial dyspraxia[76,78], and it is found in a region notably depleted in introgressed Neanderthal and Denisovan DNA[93,94].

While the small TF data set is underpowered for robust inference, the cell type sensitivity of FitCons2 provides a novel ability to distinguish between regulatory activity in different cellular developmental stages. Tissue differential analysis using FitCons2 suggest BAF155 and FOXP2 bind to the Coordinator motif at lower numbered positions and play a role in the mechanics of human craniofacial development.



Rank	TF Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Max
1	MAX	1	2	2	2	2	1	1		2	2	116	116	127	127	126	126	14	127
2	**TCF12**	1	1	1	1	1					62	114	114	114	114	114	114	60	114
3	**RAD21**	6	5	3	2	3	3	52	64	64	75	75	75	76	76	78	78	83	83
4	CTCF	1	1	1	1	2	2	40	44	44	48	48	48	49	50	50	50	52	52
5	ZNF143	4	3	3	2	3	3	35	38	38	41	41	41	42	43	43	43	45	45
6	RFX5	2	1	1		2	2	33	37	37	41	41	41	42	42	42	42	44	44
7	CMYC	2	2	2	2	2						36	36	41	41	41	41	5	41
8	**USF1**	5	5	5	5	2	1	2	2	18	17	28	28	32	32	32	34	22	34
9	**USF2**	2	1	1	1					25	25	28	28	31	31	31	32	10	32
10	WHIP					2	2	25	26	26	27	27	27	27	28	29	29	29	29
11	GTF2F1	2	1	2	2	2	5	23	23	23	25	24	24	25	25	25	25	26	26
12	STAT1	2	2	2	1	2	2	17	18	18	21	21	21	21	21	21	21	21	21
13	YY1	20	19	19	19	18	18	18	18	18	18	7	5	5	5	5	13	13	20
14	GATA2	17	16	15	18	16	15	13	4	4	3					2	2	3	18
15	BRCA1	2	2	2	1	1	13	14	13	14	14	14	14	14	15	15	15	17	17
16	CEBPB	17	14	12	12	4	4	4	4	2				1	7	7	10	10	17
17	**FOXA1**	15	14	16	16	16	12	9	9	9	7	2	2	2	2	6	6	9	16
18	NRSF		2	2	2	3	3	3	3	3	3	3	1	14	14	15	16	16	16
19	**FOXA2**	14	13	15	15	15	13	10	10	10	8	2	2	2	2	9	9	12	15
20	CHD2	10	8	6	5	5	4	3	3	5	4	4	4	4	6	9	13	13	13
⋮	⋮																		⋮
28	CJUN	7	6	7	6	6	6	3	2	2	2	1			1	2	3	4	7
29	**FOXP2**	1	1	1	4	4	7	6	6	6	6	3	3			3	3	3	7
30	CFOS	7	5	6	4	4	4	3	2	1	1				1	1	1	2	7
⋮	⋮																		⋮
33	BHLHE40										1	6	6	6	6	6	6	5	6
34	**JUND**	5	3	5	5	5	5	5	5	2	2	2			2	2	3	4	5
35	GATA1	5	5	5	5	5	5	5											5
⋮	⋮																		⋮
56	ETS1	1	2	2	2	2	2	1	1	1	1	1	1	1			1	1	2
57	**BAF155**	1	1	1	2	2	2	2	1	1	1	1							2
58	FOSL1	2	2	2	2	2	2	2											2

Rank	TF Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Max
1	MAX																		
2	**TCF12**											2	4	4	4	4	4	4	4
3	**RAD21**							1	1	1	1	1	1	1	1	2	2	2	2
4	CTCF																		
5	ZNF143																		
6	RFX5																		
7	CMYC																		
8	**USF1**	1	1	1	1					2	2	2	2	2	2	2	2		2
9	**USF2**									2	2	2	2	2	2	2	2		2
10	WHIP																		
11	GTF2F1																		
12	STAT1																		
13	YY1																		
14	GATA2																		
15	BRCA1																		
16	CEBPB																		
17	**FOXA1**	1	1	2	2	2	2	2	2	2	2	1	1						2
18	NRSF																		
19	**FOXA2**	1	1	2	2	2	2	2	2	2	2	1	1						2
20	CHD2																		
⋮	⋮																		
28	CJUN																		
29	**FOXP2**				1	1	2	2	2	2	2	1	1			1	1	1	2
30	CFOS																		
⋮	⋮																		
33	BHLHE40																		
34	**JUND**	2	2	2	2	2	2	2	2						1	1	1	1	2
35	GATA1																		
⋮	⋮																		
56	ETS1																		
57	**BAF155**	1	1	1	2	2	2	2	1	1	1	1							2
58	FOSL1																		

Figure 2.10: Transcription factor binding in enhancer Coordinator motifs. The number of times that each bound transcription factor (*TF*, by row) overlaps a Coordinator motif position (columns). The left panel, in blue, represents the complete set of coordinator motifs identified in 13,480 craniofacial enhancers. The right panel is a subset of 1,348 differential enhancers having the greatest FitCons2 score difference between a fetal tissue and an adult neural or chondrocytic cell type. A small fraction of TF hits on the left, also show up on the right. In particular, none of the highest-ranking (MAX) TFs show up on the right (outlined row, top). However, all hits associated with TF BAF155 show up in the differential set (outlined row, bottom). FOXP2, a protein associated with speech and language disorders, is found to be 3x enriched in developmental enhancers (outlined row, center).

2.2.9 Quantifying types of selective pressure in humans

As FitCons2 scores are explicitly a probability of being under selective pressure, we can estimate the fraction of autosomal sites under selection in humans by simply averaging scores across genomic positions and cell-types. The FitCons2 tree decomposition progressively divides the human genome to generate increasingly coherent subsets of positions according to human selective pressure. More coherent selective pressures lead to improved power to detect conservation and commensurately, an expected increase in the fraction of detected positions under selection. The expectation of sites under selection under INSIGHT, from all genomic positions taken together is 7.28%. This expectation increases with FitCons2 tree refinement to a maximum of 8.30%, before beginning to decline (Figure 2.21). For the present analysis, a conservative minimum of 50 bits of NLL gain was used as a termination condition, resulting in an expected value of 8.17%.

At the same time, violations of our modeling assumptions may bias INSIGHT-ρ values upwards, especially for functional classes with true scores closer to zero. The reason for this bias is that the estimate for ρ is bound at zero, so sampling noise and model misspecification can result in estimates above zero, but never in estimates below zero. To estimate the magnitude of this bias, we estimated the fraction of sites under selection, according to INSIGHT, in a collection of positions strongly depleted for known classes of mammalian conservation and genomic function. This estimate can be considered an upper bound on the misspecification error. Beginning with the collection of 1,333,467,128 function depleted autosomal positions identified in our previous work[66], we further removed potentially active positions associated with any non-null FitCons2 class in 111 karyotype normal Roadmap cell types, leaving 259,512,926 positions expected to contain very few sites under selection. Our estimate of the expected fraction of sites under selection for this collection of sites is 2.50%.

Thus, a lower bound for the fraction of sites under selection is 7.28%-2.50%=4.78%, with an upper bound of 8.30%. Our estimated range of 4.78%-8.30% for the fraction of sites under selection is fairly consistent with a variety of previous estimates, but slightly higher than the range reported for fitCons, apparently because our new model fits the data better. The INSIGHT model also estimates the fraction of constrained sites that are under weak and adaptive selective pressures. A position-wise average of weak selection yields genome wide expectations of 280.7 sites/million bp (Mbp) under weak selection, and 46.3 sites/Mbp under adaptive selection.

These FitCons2 aggregates show elevated weak selection at promoters (1,824 sites/Mbp) and enhancers (374 sites/Mbp) above the whole genome levels. Also, promoters show a 4.9-fold enrichment in sites under weak selection relative to enhancers. Surprisingly, FitCons2 classes associated with promoters (27.2 Mbp) also show a 12.9-fold enrichment in sites under adaptive selection (588 sites/Mbp) over sites associated with enhancers (46 sites/Mbp). Estimates of adaptive selection in enhancers is similar to genome wide averages. Despite substantial uncertainty in these estimates, the enriched density of weak and adaptively selected sites at promoters suggest that promoters may be a greater driver of phenotypic variation than enhancers in hominins over the past 4-6 million years.

2.3 DISCUSSION

The central idea behind the FitCons approach is that intelligible combinations of functional genomic properties characterize classes of genomic function that are predictive of selective pressure across cell-types. The functional classes represent latent categories of genomic elements that are similar to promoters or enhancers, but have been more poorly characterized. This relationship between primary sequence conservation and molecular phenotype is motivated by the observation that

biochemistry serves as a direct intermediary between primary sequence and organismal phenotypes that are subject to differential procreative fitness. FitCons2 uncovers these latent categories using selective pressure as the quantification of a generalized idea of genomic function, while the selected genomic properties provide a more qualitative and interpretable description of local genomic biochemistry.

The primary focus of the present FitCons2 implementation is to uncover latent functional classes, principally in the challenging and poorly understood noncoding genome. Selective pressure allocates 47 of the 61 FitCons2 classes to noncoding DNA (estimated to contain > 91% of conserved sites) and only 14 to human CDS. Furthermore, noncoding regulators of transcription are more plausibly under weak selective pressure than coding regions, many of which are strongly conserved and shared among a diverse range of organisms from human to yeast. Weak selective pressure can influence phenotype but must also be non-lethal in order to be observed in the population. The identification of selective constraint as central, rather than just one many of undifferentiated genomic properties, provides a grounding objective for methods like CADD, LINSIGHT and FitCons and serves to differentiate them from unsupervised data categorization techniques like Eigen or ChromHMM.

While not specifically trained to identify known classes of *cis*-regulatory elements, FitCons2 demonstrates excellent predictive properties for these elements. FitCons2 equals or outperforms not only our previously published fitCons method, but also recently developed methods like FunSeq2 and Eigen. The improvement in diversity of genomic properties used as covariates, as well as variety of cell-types utilized, provides greater coverage of types of genomic activity (such as splicing) and genomic positions involved cell-type specific activity. This expansion in properties also improves genomic resolution via precise covariates like transcription factor binding motifs. Indeed, by incorporating covariates with single base-pair resolution,

FitCons2 develops scores comparable in accuracy to systems designed to identify pathogenic single nucleotide variation. The straightforward clustering scheme for covariates produces a generative joint probability distribution over genomic states and selective pressure that is not only highly predictive of genomic function, but also readily interpretable and highly scalable to additional cell-types, cell-states, and covariates.

2.3.1 Intelligibility, generative modeling and computational challenges

In addition to predictive accuracy, this research accomplishes a second objective in terms of intelligibility. Each inferred FitCons2 functional class, is constrained to be a straightforward combination of genomic properties represented in quantized levels of covariate activity such as {High, Medium, Low, None}. The functional classes are informative across all investigated cell types leading to small set of 61 classes that are readily interpretable and universal to a set of cells representative of a broad range of human tissue types (see section 2.5.1 Cell-type specific regulatory activity). Machine learning systems such as Neural Networks (DANN[95]), and generalized linear models (CADD, LINSIGHT, Eigen) create diffuse models with information spread over thousands of parameters, each contributing a small amount to a prediction. While the flexibility of such distributed models can provide additional prediction accuracy, it deeply obscures scientific understanding. Furthermore, such diffuse models are typically discriminative, that is they do not model relationships among the covariates and cannot be used to make predictions if any covariate value is missing. FitCons2, in contrast, is generative. The complete distribution over quantized covariate values is known and can be used to estimate class segmentation and scoring on novel cell types, even if some of the covariates are uncertain, missing or unmeasurable in the novel cell type. In essence, discriminative models can be thought of as the resolving a single specific question (like classing HGMD variants), while

generative models can be thought of as providing insight into a board space of problems in the relationship between functional and selective genomic properties. As a tool for researchers, existing FitCons2 scores can be used to provide comprehensible answer questions that scientists have not yet thought to ask.

2.3.2 Selective pressure on epigenetic marks

In FitCons2 we utilize selective pressure to infer optimally informative patterns of genomic properties, and jointly assign each such pattern a score based on selective pressure. We approach this joint inference by considering the expectation of selective pressure over a collection of genomic positions demonstrating a particular pattern of genomic properties in a particular cell type (a relatively intuitive concept). However, to be admissible we also require that a pattern be informative across **all** cell-types in a training set. As a particular pattern of genomic properties may occur at differing positions in differing cell-types, aggregating across cell-types attempts to integrate out the cell-type specific genomic position as a nuisance parameter. This aggregation leaves a direct association between genomic properties and expected selective pressure as measured by ρ. The resultant measure of selective pressure can be directly interpreted as the probability that a genomic position having a particular pattern of properties in a karyotype normal cell-type, is under selective pressure. Under the assumption that actual selective pressure is associated with a generalized concept of genomic function, a FitCons2 score can be interpreted as the potential for genomic function in any position evincing a given pattern of genomic properties. Such a functional potential might, for example, be applied to identify pathogenic dysregulation in affected tissue (such as the alternative promoters the develop in oncogenic processes) even though such promoters may never come under direct selective pressure.

2.3.3 Informing experimental design

The FitCons2 model quantifies information about selective pressure in terms of likelihood, which can be interpreted as bits of information about selective pressure. By considering selective pressure as a measure of generalized genomic function, the FitCons2 framework can answer experimental design questions simply and quantitatively in terms of the expected information gain about genomic function from a proposed experimental assay. In considering expected information gain, it is particularly important to recognize that properties that are highly informative individually may have low additional information content conditioned on a preexisting set of measurements. Conversely, carefully selected measurements may have modest individual informative content, but render previously observed variables **more** informative in a process called informational synergy[96]. FitCons2 provides estimates of the experimental effectiveness of genomic measurements (see Figure 2.11), which can in principle be translated to cost per bit of expected information gain to prioritize experiments. More information about the contribution of each covariate in the actual FitCons2 model, as well as the information theoretic impact of removing each covariate can be found in supplemental section 2.5.5 Information theoretic properties of covariates.

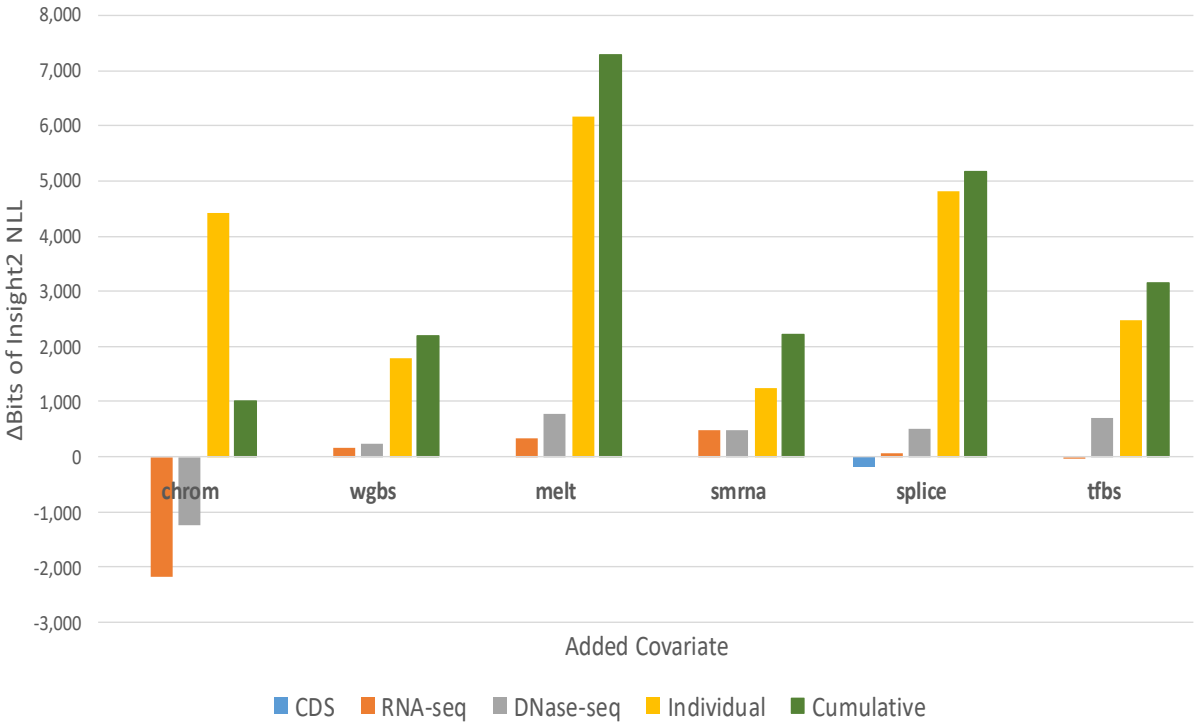


Figure 2.11: Information impact of added covariates. The results of six FitCons2 tree decompositions, each with four covariates: three shared among all runs (CDS, RNA-seq and DNase-seq), and a fourth added to identify its impact on other covariates and total model negative log likelihood (*NLL*, in bits). Along the vertical axis we see the difference in NLL scores. Along the horizonal axis are a set of bars for each added covariate. The height of each bar shows an NLL attributed to splits in a FitCons2 tree. The first three bars represent the three shared covariates (CDS in blue, RNA-seq in orange and DNase-seq in grey). The fourth bar shows information attributed to the added covariate (yellow). The sum of these first four bars is represented as a cumulative bar, in green. Net impact of an added covariate is always positive. The blue bar is often too small to be visible, but is noticeable in the "splice" set. The "chrom" (chromatin state) covariate shows *redundancy* by individually adding 4,425 bits to the model but simultaneously reducing contributions of DNase-seq and RNA-seq so that the net impact on the model is only 1,006 bits. The smRNA (small RNA-seq) covariate shows *synergy*, providing only 1,255 bits of information individually, but improving the information provided by of RNA-seq and DNase-seq by 479 and 484 bits, respectively, for a total impact of 2,218 bits.

2.3.4 FitCons as an extensible framework

The FitCons2 framework involves identifying patterns of functional covariates that maximally informative about genome-wide selective pressure. The present investigation constrained investigation to pattern search based on recursive bipartition, and selective pressure measure based on INSIGHT. However, a wide variety of methods could be employed in either of these steps trading off intelligibility, precision, evolutionary time frame, and computational efficiency. In the early phases of this work less robust estimates for hominin selective pressure were investigated and results were found generally compatible. Selective pressure was chosen as representative of generalized genomic function, to identify a diverse subset of functional classes. If the objective of interest is more focused, for example the probability of association with a particular phenotype or even disease process, the FitCons2 methodology can similarly be used to identify functional classes most relevant to that phenotype or process.

2.3.5 Functional classes as a lexicon for developmental regulatory activity

The collection of 61 classes identified by FitCons2 forms a basis for the representation of functional genomic state in terms molecular phenotypes identified by selective pressure. A natural extension would investigate the spatial dynamics of state transitions across the genome for an individual cell state. A spatial classification of patterns could be approached using hidden Markov models or Bayesian nets. In turn, a spatial classification of patterns would simplify characterization of temporal dynamics of normal cell development and disease, as well as suggesting a structure for long range interactions via correlated state transitions across potentially wide genomic spans. Thus, FitCons2 classes can be considered a step towards a selection-derived basis for a description of healthy cell development derived from genomic features.

Deviations from normal development may signal the earliest signs of pathogenic processes, even before accumulated dysregulation produces a clear disease phenotype.

2.4 METHODS SUMMARY

This Methods section is typically provided in the online full-text version of a Nature style publication, but not in the print version.

2.4.1 Covariate generation

Covariates were retrieved and, as necessary, transformed into to a quantized representation providing discrete covariate value for each position in each cell-type. Thus, each covariate represents a complete partition over autosomal hg19 genomic coordinates. Of the 9 covariates, 4 were available for each cell-type and referred to as *CTS* (cell type sensitive), 3 were properties of individual positions insensitive to cell type and called *Annotations*, and the remaining 2 (TFBS and smRNA-seq) were available only for a limited set of cell-types. Data sets available for limited cell types were aggregated into a set of *Pseudoannotations* and treated as annotations in subsequent calculations.

2.4.2 Functional genomic data

Imputed *RNA-seq*, *DNase-seq*, *WGBS* and *Chromatin State* data sets for each of 127 cell types were available from the Roadmap Epigenomic Project[59]. Methylation state identified from whole genome bisulfide sequencing (WGBS) was binarized into hypomethylated and non- hypomethylated regions using the HMR program from the MethPipe package provided by the Smith lab at USC[97]. RNA-seq and DNase-seq were quantized into 4 levels each using a maximum likelihood method based on an exhaustive monotonic search over partition boundaries, with likelihoods based on Shannon entropy of INSIGHT-ρ for each continuous range of values as per fitCons, see 1.6.3.2 Mutual Information and Conditional Entropy.

2.4.3 Annotation preparation

Protein coding regions (CDS) and splice site distances were identified using data from the GENCODE V19[55] database of protein coding transcripts. A conservative subset of complete known transcripts was used to identify CDS covariates, but a less conservative superset of potential exons was also generated and used as a filter for determining confident noncoding regions in statistical tests (as per 1.5.5 GENCODE annotations). The CDS covariate was broken into 5 values consisting of: start codon, codon position 1, codon position 2, codon position 3, and Noncoding positions. A position belonging to more than one class in differing protein isoforms was assigned to the class under greatest selective pressure. The splice distance annotation was constructed by taking a metaplot of all genomic positions 50 bp upstream and downstream from each CDS exon/intron splicing junction. This generated approximately 100,000 genomic positions for each integral distance from a splice boundary. Donor (5') and acceptor (3') splice sites were treated separately. The relationship between selective pressure as measured by INSIGHT-ρ, and distance from splice site was found to be nonmonotonic. Intronic positions within 50 bp of splice sites were classified into 4 values {Hi, Med, Low, None} based on the level of selective pressure found at each (see 2.5.2.8 Splicing annotation). Of exonic positions within 50 bp of a splice site, only the two positions immediately adjacent to a splice site were found to have selective pressure in excess of nearby selective pressure at CDS. These two positions were added to the most highly conserved splice class.

Melting temperature was also identified as an important annotation that correlated highly with local GC nucleotide density. Melting temperature provides information as to the amount of energy needed to separate strands of DNA as a preparatory step for a variety of biochemical interactions, including transcription. A metaplot of INSIGHT-ρ versus melting at 0.5°C intervals indicated a relatively

smooth structure with elevated selective pressure at very high and very low melting temperatures and minimal selective pressure at central melting temperatures. Melting temperature was quantized into 5 levels {VeryLow, Low, Medium, High, VeryHigh} and designated as a nonmonotonic covariate for FitCons2 decision tree inference. For details, see 2.5.2.9 Melting temperature annotation.

2.4.4 Pseudoannotations

Both Transcription Factor Binding Sites (*TFBS*) and small RNA-seq data sets were available for a very limited collection of cell types and Transcription Factors (*TF*s). However, both of these data sets were considered to have high biological relevance. TFBS are central to the activity of enhancers and promoters, while small RNA-seq provides coverage of biologically active micro RNAs that are not necessarily covered by the conventional RNA-seq measurements. To utilize the available small RNA-seq data sets, small RNA-seq measurements were aggregated across cell-types into a single covariate indicating the potential for small RNA transcription. This aggregate is called a *pseudoannotation*. Pseudoannotations are applied equally to all cell-types in the FitCons2 tree decomposition where they are treated like regular annotations such as CDS. In FitCons2, actual cell-type specific activity associated with a pseudoannotation is inferred through conditional combinations of cell-type sensitive covariates such as DNase-seq with a pseudoannotation such as TFBS.

Transcription Factors (*TF*s), and TF binding motifs were available for a limited set of transcripts from and cell types from the Ensembl Regulatory build V75[60] and from Leo Arbiza[28]. Both sets were developed by their authors from ChIP-seq data via joint inference of binding motifs (as position weighted matrix[77], *PWM*) and positions. The Arbiza set was found to be larger with better recall, while the Ensembl set smaller and more precise. The information content of each PWM position at each genomic

binding site was identified and use as a monotonic covariate for maximum likelihood quantization. In the case that a genomic position belonged to more than one PWM position, the PWM providing a higher information score was used for that genomic position. The Ensembl and Arbiza sets were quantized independently into 4 classes, and the resulting combination of 16 class pairs again quantized into the 4 classes of the TFBS covariate using the same ML method used for continuous covariates (see 2.5.2.10 Transcription factor binding site pseudoannotation).

Small RNA-seq (*smRNA*) was obtained in two sets, the first was the UCSF-4Star composite and the UCSF Brain Germinal Matrix. These two data sets were cell composites and were quantized independently into 3 levels. Each set provided a partition on genomic positions. The cross product of these two partitions resulted in a new genomic partition with 3x3=9 classes and was requantized into 4 levels. Three other data were included, two from ENCODE cell lines (CD20 and HUVEC) and one UCSF Penis Foreskin Keratinocyte composites (*PFK*). These data sets were combined in a manner similar to the initial two sets. In each of these cases, replicates were quantized to three levels and then the cross section requantized into three levels. This 3x3x3=27 level set was requantized into 4 levels. Finally, the two larger data sets were requantized from 4x4=16 classes into the 4 class. For details, see 2.5.2.11 Small RNA-seq pseudoannotation.

2.4.5 INSIGHT optimization

The INSIGHT program[34] is a species agnostic selective pressure inference engine that is designed for flexibility, but not computational speed in measuring human selective pressure. INSIGHT database input-output (IO) was reimplemented for rapid binary access to human-specific data and its expectation maximization (*EM*) based parameter inference was replaced with a faster bounded gradient descent method over negative log likelihood (*NLL*). The software's features were extended to:

allow for weighting of individual genomic positions (as a fraction of all cell types), allow for priors to be added in each calculation as a number of pseudocounts from a distribution defined by user-provided parameters ρ, η, and γ. Also, the INSIGHT databases containing divergence (λ) and polymorphism (θ) block estimates revised, with priors replacing heuristic cutoffs for data quality management. The result was a program that ran 9,960x faster and was less susceptible to convergence failure for small data sets, but was specialized for human selective pressure inference. The polymorphism database, primate divergence references and neutral sites filter were all unmodified from the original INSIGHT[34].

2.4.6 FitCons2 decision tree training

Roadmap cell types are numbered E001-E129, with two numbers (E060, E064) missing. Of the 127 remaining cell lines, 5 were removed as being annotated karyotype abnormal (E114, E115, E117, E118, E123) and 7 were removed for data quality deficit annotations (E001, E003, E017, E027, E098, E104, E113). The remaining 115 were used as a training set for FitCons2. After training, scores and segmentations were calculated for all 127 cell types. Karyotype abnormal cells were removed in an attempt to sample more directly from the sorts of epigenomic signals that might be indicative of sites under selective pressure in healthy tissue. The first FitCons2 tree split (at the root) took 4 hours and 52 minutes to complete on 4 CPUs using less than 32 GB memory. As each split of a tree node is a partition, the number of positions considered at each tree **depth** is constant. Thus, the first split is a reasonable estimate for the total CPU time required to calculate all splits at a single tree depth. All internal nodes at a given tree depth can be executed in an "embarrassingly-parallel" fashion. The resultant tree had a minimum leaf depth of 5, and a maximum leaf depth of 12, before reaching the chosen cutoff of 50 bits (34.8 nats), this represents a LRT cutoff of 69.6 nats, where each new split adds

approximately 4 degrees of freedom to the model, consisting of the split location and the 3 additional INSIGHT parameters developed at the leaves. Training completed in under 2 days and 9 hours of wall time on a large computational cluster.

2.4.7 Cell-type independent score generation

To generate an aggregate scoring of each genomic position, across cell-types, we wanted to: maintain the interpretation of score a selective pressure, maintain sensitivity to tissue and developmentally specific activity, and compensate for a potentially unbalanced selection of tissue types. To accomplish this, we generated a numerical weight for each cell type representing its informative content, and then combined weights and scores across cell types at each position to generate a single scoring for that position. In the last step, the FitCons2 tree decomposition method was again used to balance distributions of weights and scores across cell-types, while greedily maximizing information about selective pressure.

To attempt to weight each cell-type according to its informative content PCA was performed with 2.88 billion genomic positions as random variables across the 115 training cell-types (as observations). The result was 114 orthogonal vectors of length 2.88 billion. The directed projection of each cell type onto each principle component (*PC*) was aggregated across components so that each unit of distance along a PC was allocated only once and shared proportionately among all cell types with non-0 projections onto the component. The use of all 114 components was necessary to preserve both the potentially unique regulatory properties of each cell type, while simultaneously distributing redundant projection proportionately among the cells that shared it. Dividing each cell's attributed weight by the mean projection length for a single vector provided a relative weight between 0.065 and 0.421 for each cell type (median 0.16). The sum of the weights was 19.98 suggesting that the 115 Roadmap cell types together carried the functional variation of approximately 20 "independent"

cell types. To generate the independent scoring 12 ρ-thresholds were selected between 0.0 and 1.0 using the quantization method described above for individual covariates. Each cutoff was treated as representing a separate covariate. Each of the 13 classes ($x \in \{1 \ldots 13\}$) was treated as a covariate C_x with a score cutoff $Cut(x) \in [0.0 - 1.0]$. At a genomic position i, the continuous covariate value $C_x(i)$ was the sum over cell-types ($ct \in \{Cell\ Types\}$) of weights $w(ct) \in \{0.065 - 0.421\}$, for cell-types with FitCons2 score ($Score(ct, i)$) greater than or equal to the covariate cutoff, $Cut(x)$. Formally,

$$C_x(i) = \sum_{ct \in celltypes} w(ct) * 1_S(ct, i, x)$$

Where

$$1_S(ct, i, x) = \begin{cases} 1 & \text{if } Score(ct, i) \geq Cut(x) \\ 0 & \text{otherwise} \end{cases}$$

Each covariate C_x was then individually quantized using standard covariate processing to 5 levels, representing covariate-specific weight thresholds. These quantized covariates were used as the basis for a second FitCons2 tree inference that generated 37 leaves at a significance cutoff of 50 bits. As the covariates are aggregated across cell-types there is no further cell type dependency in the score (*CTI* scores). The score at each leaf is the result of an INSIGHT-ρ estimate so interpretability as fraction of sites under selection remains intact and directly comparable with cell-type specific (*CTS*) scores. This method neither suffers from asymptotic aggregation of noise (as taking the max over scores would), nor depletion of score precision (as taking the average would). The 37 CTI scores range from 0.0381 to 0.8844 with 8 values < .1 and 22 values < .2.

2.5 APPENDIX II / SUPPLEMENT TO SECOND PAPER

The following material would not be part of the printed matter in a Nature-formatted version of the paper, nor would it appear in the full-text online version. However, this material is typically provided by the publisher as an associated resource that is linked to the online version of the paper. In the traditional dissertation format, such material would likely be provided in an Appendix at the end of the document. However, in a papers-format dissertation, guidelines provided by Cornell University request all material from a publication be maintained in the same dissertation chapter. To comply with Cornell University guidelines, relevant appendix material is provided in the following section.

2.5.1 Cell-type specific regulatory activity

Figure 2.12: Comparison of EGFLAM promoter activity in H1-hESC and GM12878. The first codon in the EGFLAM gene shows strong transcription in H1 hESC, but little or no transcription in GM12878 (green covariate, H1 and GM). Corresponding FitCons2 scores (dark blue) over the region 200bp upstream of the transcription start site are higher in H1 hESC then in GM12878. Transcription proceeds from right to left. The primary FitCons2 Class for H1 is 22 (dark purple), an active promoter class with a score of 0.21, while the primary FitCons2 (FC2) class for GM is 60 (mauve) a weakly active intergenic class with a score of 0.03. Positions associated with factor binding (TFBS row, gold) are highlighted vertically in yellow. These positions in H1 are in classes 16 or 19 indicating active binding with a score of 0.43, while the same positions in GM are in lower scoring class 44 (less active TF, score 0.40) or 39 (non-TF intergenic, score 0.33). While less obvious, codons in the exon body of H1 have a minimum score from class 10 (score 0.63) versus the same positions in GM which fall into class 11 (score 0.55). The active exon splicing boundary, highlighted vertically in red, shows spikes in H1 with class 09 (Active Splice, score 0.87) versus GM's class 12 (Untranscribed CDS, score 0.42) at the same position. Functional scores LINSIGHT and phyloP (bottom rows) both show signal peaks in highlighted areas, but are unable to identify cell-type specific differences in activity.

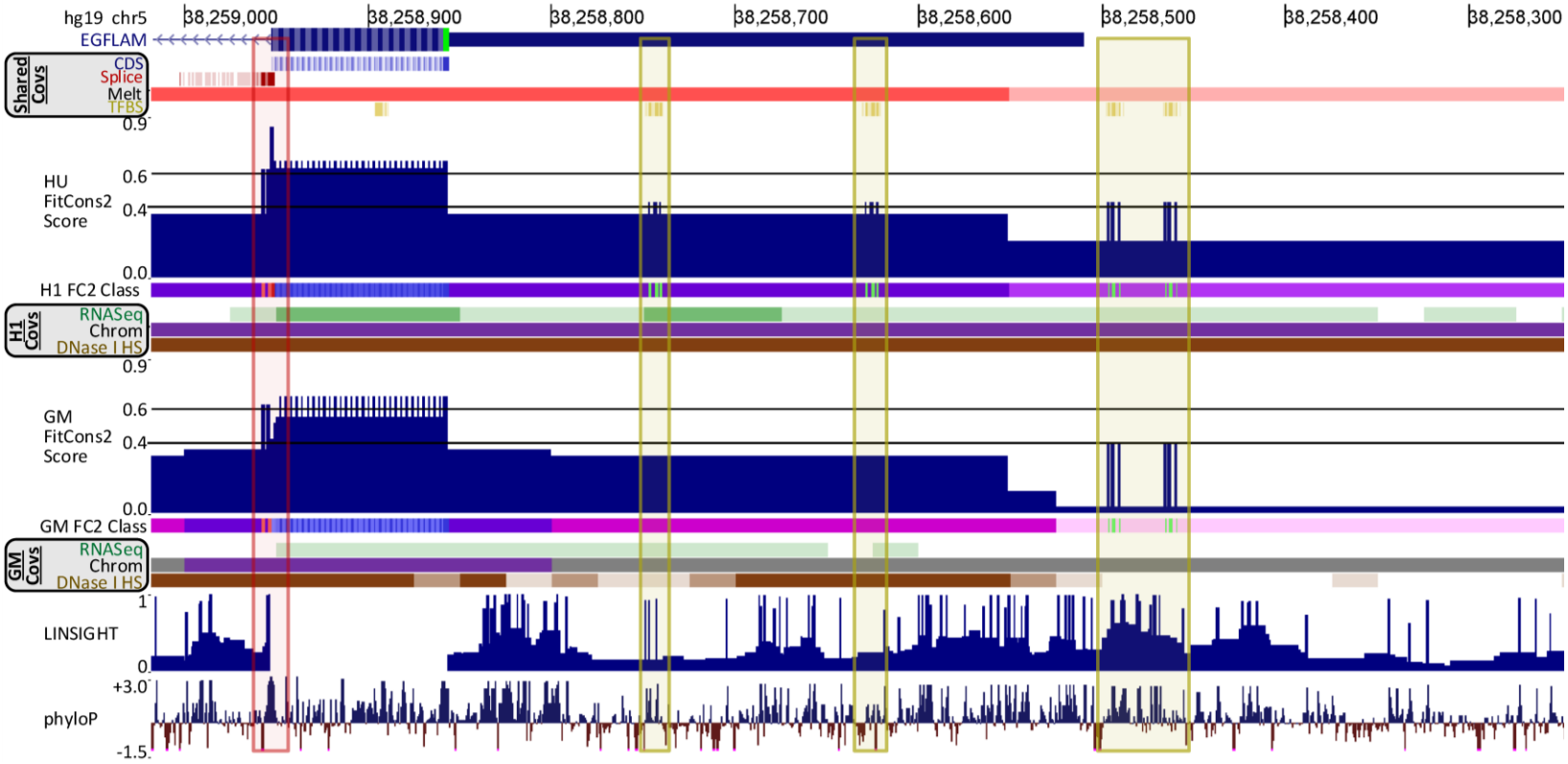

hg19 chr5
38,259,000
38,258,900
38,258,800
38,258,700
38,258,600
38,258,500
38,258,400
38,258,300
EGFLAM
Shared Covs
CDS
Splice
Melt
TFBS
0.9
HU FitCons2 Score
0.6
0.4
0.0
H1 FC2 Class
H1 Covs
RNASeq
Chrom
DNase I HS
0.9
GM FitCons2 Score
0.6
0.4
0.0
GM FC2 Class
GM Covs
RNASeq
Chrom
DNase I HS
1
LINSIGHT
0
+3.0
phyloP
-1.5

Figure 2.13: Comparison of LCT1 super-enhancer activity across three cell types. A super-enhancer is considered a proximal collection of enhancers that contribute to the regulation of the same gene. The super-enhancer SE33394 (above) (Khan and Zhang, 2016) is more than 2 Mbp away from its associated gene LCT1 (below). In H1, LCT1 shows both strong RNA-seq signal and chromatin state associated with transcription (A), while showing neither in GM (B) or HU (C). Highlighted in gold, the super-enhancer has 5 loci designated distal regulatory modules (blue, (Gerstein et al., 2012)) as well as a FANTOM5 enhancer (green, (Lizio et al., 2015)), and is flanked by two GWAS hits associated with observable phenotypes (highlighted in red, rs10507601 and rs9527419) . The active enhancer in H1 has ambient score plateaus of class 20 (score 0.45) and class 27 (score 0.27) and with peaks in regions of small RNA-seq signals to class 26 (score 0.41). In inactive cell types, the class is typically 56, 58 or 48 all with scores below 0.09, and occasional peaks in inactive transcription factor binding sites with class 45 (score 0.31). Neither LINSIGHT nor phyloP find elevated scores in the left ⅔ of the enhancer locus. While conservation scores are elevated in the right ⅓, the primary annotation signature there is decreased melting temperature. Decreased melting temperature is broadly indicative of increased selective pressure but absent other functional signals it is only a weak driver of elevated FitCons2 scores.


hg19 chr13
56,163,000 56,162,000 56,161,000 56,160,000 56,159,000 56,158,000 56,157,000 56,156,000 56,155,000 56,154,000 56,153,000 56,152,000 56,151,000 56,150,000 56,149,000 56,148,000 56,147,000 56,146,000
DRM
FANTOM
GWAS
Shared Covs
TFBS
smRNA
Melt
A) ACTIVE
H1 FitCons2 Score
H1 FC2 Class
H1 Covs
RNASeq
Chrom
DNase I HS
HypoMeth
B) INACTIVE
GM FitCons2 Score
GM FC2 Class
GM Covs
C) INACTIVE
HU FitCons2 Score
HU FC2 Class
HU Covs
LINSIGHT
phyloP
hg19 chr13
53,310,000 53,305,000 53,300,000 53,295,000 53,290,000 53,285,000
LCT1
A
RNASeq
Chrom
B
RNASeq
Chrom
C
RNASeq
Chrom

2.5.2 Covariate Development

Below is a summary of the genomic properties used as covariates in FitCons2. Following sections detail the source and development of each. Covariates fall onto three types: cell-type sensitive (*CTS*), annotations, and pseudoannotations. CTS data, such as RNA-seq, is available separately for each position in each cell-type, and can vary by cell-type. Annotations represent genomic properties that are fixed and therefore shared among all cell-types. Examples of annotations include protein coding (*CDS*) phase and splice site distance. Pseudoannotations represent CTS data sets that are available for only a small number of cell-types. Pseudoannotation such as small RNA-seq is aggregated into a set representing the potential for covariate activity at each genomic position, and then applied as an annotation to all cell-types.

In addition to covariate type, covariates also are divided by monotonicity. A monotonic covariate is considered to have a greater impact on selective pressure with increasing class quanta. For example, a position in the monotonic RNA-seq class 4 has more transcription activity than a position in RNA-seq class 3. It is assumed that higher numbered monotonic classes have a nondecreasing magnitude of impact on selective pressure, either uniformly increasing selective pressure, or uniformly decreasing it. This assumption limits the model complexity when performing exhaustive searches over partitions of one covariate to $O(N)$, when there are N levels in a covariate. Nonmonotonic covariates classes (DNA melting temperature and chromatin state), have an unknown ordering with regards to selective pressure. Before each new split in the decision tree, nonmonotonic covariates are sorted via conditional INSIGHT-ρ values. This is accomplished by a conditional partitioning of all remaining genomic positions according to the nonmonotonic covariate class values. Once ρ is calculated for each covariate class value, the class values are sorted by ρ and this new ordering of classes treated as a monotonic covariate for the current split.

Only positions in the node under consideration are included in this reordering, with this limitation forming the basis for the "conditioning". The ordering of classes in nonmonotonic covariates can change at each node in the fitCons decision tree.

Preliminary quantization studies for cell-type specific data were performed over a subset of cell-types for which both raw and imputed data were available. This availability varied among covariates.

Table 2.1: FitCons2 covariate summary. Listing of the nine FitCons2 covariates including covariate Name, Covariate Type (Annotation, Pseudo Annotation or CTS), number of quantized Levels for each covariate along with level mnemonics, Source for each covariate, and property of Monotonicity indicating weather increasing levels as seen as having a progressively greater impact on selection in a uniform direction. Nonmonotonic covariates are reordered (according to marginal INSIGHT score) before each before each split is calculated. Marginal information in megabits is estimated from 14 sample cell types and is information about the indicator variable for selection at each genomic position $S \in \{0,1\}$, where $\rho = [S]$. This measure is useful for judging the relative informativeness of covariates, but is on a different scale from INSIGHT-NLL values described in the FitCons2 decision tree.

Name	Type	Levels	Source	Mono?	Marginal Inf. Mbits
CDS	Annotation	5: Start, phase0,1,2, NCD	Gencode19	Y	48.1
Splice	Annotation	4: Core, Prox, Dist, Non	Gencode19	Y	5.5
Melt	Annotation	5: VHi, Hi, Med, Lo, VLo	MeltMap[98]	N	40.7
TFBS	Pseudo Annotation	4: Hi, Med, Lo, None	Ensembl, Arbiza et al	Y	5.4
smRNA	Pseudo Annotation	4: Hi, Med, Lo, None	4Star, BGM, PFK, HUVEC, CD20 cells.	Y	19.7
RNA-seq	CTS	4: Hi, Med, Lo, None	Roadmap	Y	40.7
DNase-seq	CTS	4: Hi, Med, Lo, None	Roadmap	Y	10.0
Chrom State	CTS	25[59]	Roadmap	N	23.6
WGBS	CTS	2: Hypo, non-Hypo	Roadmap	Y	4.1

2.5.2.1 General quantization method

Quantization for continuously valued covariates proceeds as per 1.6.3.4 Application to RNA-seq data. In summary, continuous values are quantized by partitioning values into fine contiguous ranges based on the number of genomic positions in each range, generally this is along percentile boundaries targeting 1 percent of genomic positions in each bin. INSIGHT ρ is calculated for each fine bin. To create a coarse quantization of this fine quantization, an exhaustive partitioning of the finely discretized value range into N classes is explored, where N is typically in the range two to six. For a given N the partitioning that produces the lowest expected conditional Shannon Information. This process is equivalent to maximizing the mutual information between the coarsely quantized classes and the INSIGHT-ρ scores of the finely quantized positions. The value of N is generally selected based on heuristic intelligibility considerations as statistical significance tests on a single covariate will generally justify undesirably large values of N.

In the FitCons2 covariate quantization information for a collection of N sites with fraction of sites under selection of ρ was taken as the Shannon Information of a latent random variable $S \in \{0,1\}$. S takes the value 1 if a position is under selection and 0, otherwise (see INSIGHT model) and ρ is an estimator for $[S]$. The expected Shannon Information can be calculated as $-N(\rho \log_2(\rho) + (1-\rho)\log_2(1-\rho))$. In the information tree decomposition, INSIGHT-NLL was used instead as the source for information about selective pressure, as this provided improved resilience to model misspecification of ρ. While strongly correlated, the number of bits attributable to INSIGHT-NLL are on a different scale from the number of bits derived from the Shannon Information about S, so bit counts generated during covariate quantization are generally several orders of magnitude larger and cannot be directly compared to bit counts generated during FitCons2 tree decomposition.

2.5.2.2 General aggregation method

When data from differing sources need to be merged into a single covariate, data from each source is quantized separately into a coarse classification via the general quantization method described in 2.5.2.1 General quantization method. This typically results in a quantization consisting of three or four classes per covariate. To combine quantized covariates, the cross product is taken, the elements of the cross product ordered according to ρ and a new partitioning of a given arity is found via exhaustive search on the ordered cross product elements, via the procedure outlined in 2.5.2.1 . Consider the example of quantized covariates X and Y having $N_X = |X|$ and $N_Y = |Y|$ unique values, respectively. Each covariate partitions the genome, that is every genomic position corresponds to exactly one value in $X = \{X_1, \dots, X_{N_X}\}$ and exactly one value in $Y = \{Y_1, \dots, Y_{N_Y}\}$. We construct a new set $Z = X \times Y$, where $N_Z = N_X N_Y$ and every position in the genome is in exactly one element of $Z = \{Z_1, \dots, Z_{N_Z}\}$. We can now calculate INSIGHT-ρ for the positions in each Z_k as $\rho(Z_k)$, then order the elements of Z according to $\rho(Z_k)$, and then find the maximally informative partitions with a coarser quantization $N'_Z \leq N_Z$ using exhaustive ordered search as described in 2.5.2.1. Generally, N'_Z is the same size as, or slightly larger than, $\max(N_X, N_Y)$. The extension to of small numbers of covariates (larger than 2) can proceed by extending the cross product $Z = X \times Y$ to the cross product of all covariates, before ordering by ρ and requantization of Z.

If there are numerous primary sources to be integrated into a single covariate, as with the small RNA-seq covariate, an alternative may be preferable. To restrict loss in power in estimating ρ when the N_Z would become large under a complete cross product, covariates can be recursively aggregated while attempting to maintain similar marginal information among combined elements. Thus, the two least informative covariates might be designated X and Y, and these two combined into a Z that replaces

the pair. This Z may be subsequently combined with other covariates with similar marginal information content. This process is similar to Huffman encoding and attempts to maintain a relatively low dimensional representation of the data over the complete cross product, while allowing differing informative properties for each combined subset.

2.5.2.3 RNA-seq

RNA-seq data was downloaded from the following location:

http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidatedImputed/RNAseq/E{XXX}-RNAseq.imputed.LogRPKM.signal.bigwig ,

where {XXX} is replaced by each of the 127 active cell-type IDs within the range 001-129. Data for each cell-type was quantized separately using the standard method, into four levels. The number four was chosen from an investigated range of 2-6 as it provided a relatively intelligible characterization as {None, Low, Medium, Hi} while simultaneously capturing most of the marginal information available.

Class	Label	ρ	# Pos	% Pos
3	Hi	0.4762	34,092,644	1.18%
2	Med	0.2146	43,806,982	1.52%
1	Low	0.0963	492,571,796	17.10%
0	None	0.0592	2,310,561,864	80.20%
	Mean→	0.072840471	2,881,033,286	**←Sum**

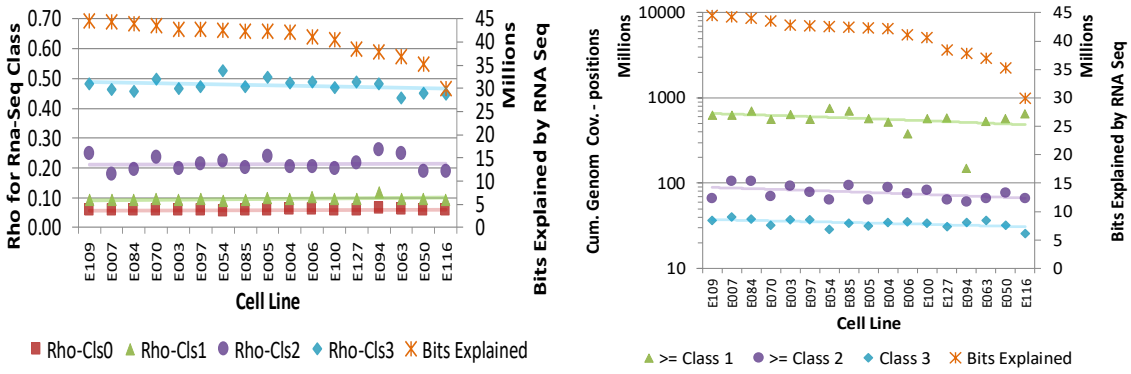


Figure 2.14: RNA-seq covariate information. The table at top provides information about RNA-seq covariate quantization including: quantized classification, label, expected fraction under selective pressure, size, and fraction of genomic positions for each value in the quantized RNA-seq covariate for 17 cell-types. Below left, a figure representing the homogeny of the covariate in terms of ρ for each class and total amount of information about selective pressure indicator $S \in \{0,1\}$, an indicator variable with estimator $\rho = [S]$. Below, right a chart showing consistency of genomic coverage and a repetition of information explained by RNA-seq in each cell type. While quantization boundaries were identified separately in each cell-type, characteristic statistics for each classification are relatively stable across cell-types.

As RNA-seq signal is relatively sparse along the genome, a signal equivalent to even a single read was generally enough to classify a position as "Low", and positions classified in the "None" class typically have 0 signal. RNA-seq data is provided as imputed, log RPKM (Reads Per Kilobase of transcript per Million mapped reads) corresponding to RNA-seq signal strength. Imputed data was used as raw RNA-seq read data was only available for 57 of the 127 cell-types and the imputation process reportedly improved on the reliability of individual raw RNA-seq data sets[99].

2.5.2.4 DNase-seq

DNase-seq data was downloaded from the following location:

http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidatedImputed/DNase/E{XXX}-DNase.imputed.pval.signal.bigwig ,

where {XXX} is replaced with the 127 valid Roadmap cell-type IDs in the range 001-129. DNase-seq data was processed much the same as RNA-seq data (see, 2.5.2.3 RNA-seq). Raw read data was only available for 53 of the 127 Roadmap cell-types. As imputed data was available for all cell-types, imputed data was used. A variety of processed data sets were available and the figure below shows how various forms of processing effect the marginal information about selective pressure provided by this covariate. As raw, imputed and peak data were all available for cell type E003 (H1 hESC) a comparison among the levels of processed data was generated. In general, increased processing removed information from single-cell data. Thus, the information levels are reduced from raw reads, to fold change (fc), to p-values of fold change. DNase broad peak and narrow peak data, which utilized an alternate processing pipeline to identify collections of contiguous positions with elevated DNase I signal was processed into 3 classes as per the fitCons covariate, and showed the lowest informative value. Only imputation, which aggregates data from multiple cell-types along with multiple covariates within the target cell-type, improved the informative value of this covariates.

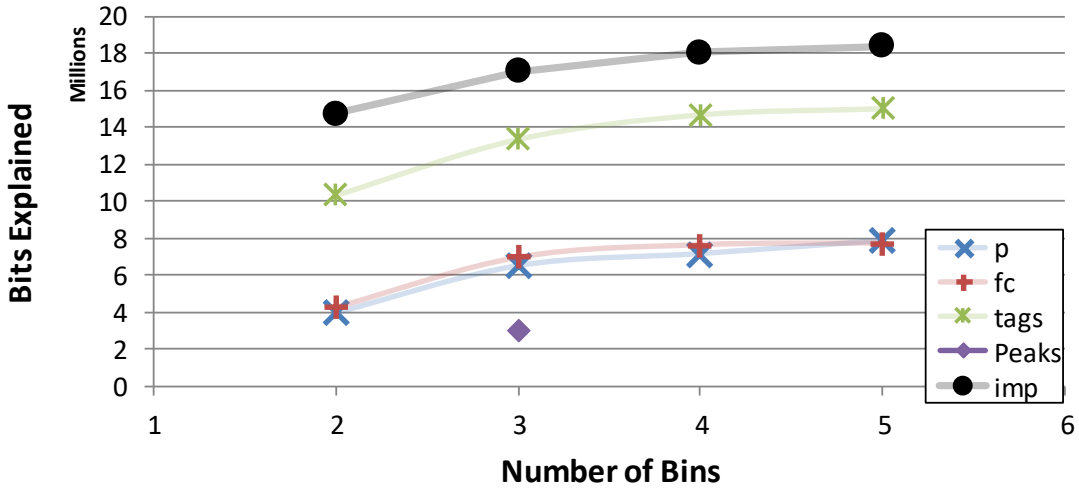


Figure 2.15: DNase-seq information content for E003, H1 hESC. Segmentation of the genome using DNase-seq provides monotonically increasing amounts of information with increasing numbers of covariate values (horizonal axis, "Bins"). However, increased statistical processing of raw tag data (green) to fold change (red), to p-value of fold change (blue) progressively reduces information at each degree of quantization. The use of a 3 category system based on the peak analysis used in fitCons {Narrow, Broad-Narrow, None} shows the lowest amount of information about selective pressure indicator variable S ($S \in \{0,1\}, \rho = [S]$). Among processed data forms, only imputed data (black) which aggregates other covariates as well as DNase-seq from other cell-types, increases marginal information provided by this covariate.

2.5.2.5 Whole genome bisulfide sequencing

Whole genome bisulfide sequencing (WGBS) data was downloaded from the following location:

http://egg2.wustl.edu/roadmap/data/byFileType/signal/consolidatedImputed/DNAMethylSBS/E{XXX}-DNAMethylSBS.imputed.FractionalMethylation.signal.bigwig ,

where {XXX} is replaced with the 127 valid Roadmap cell-type IDs in the range 001-129. This data is imputed methylation fraction at each CpG site. Methylation is a chemical modification to DNA that is associated with gene silencing. In human DNA, most potentially methylated sites are methylated. Depletions in

methylation (hypomethylation) are measured via the whole genome bisulfide sequencing (WGBS) assay. Several attempts were made to classify raw WGBS signals into levels, and none provided substantially higher marginal information content than simply bipartitioning of genomic positions for each cell-type into hypomethylated and non-hypomethylated regions.

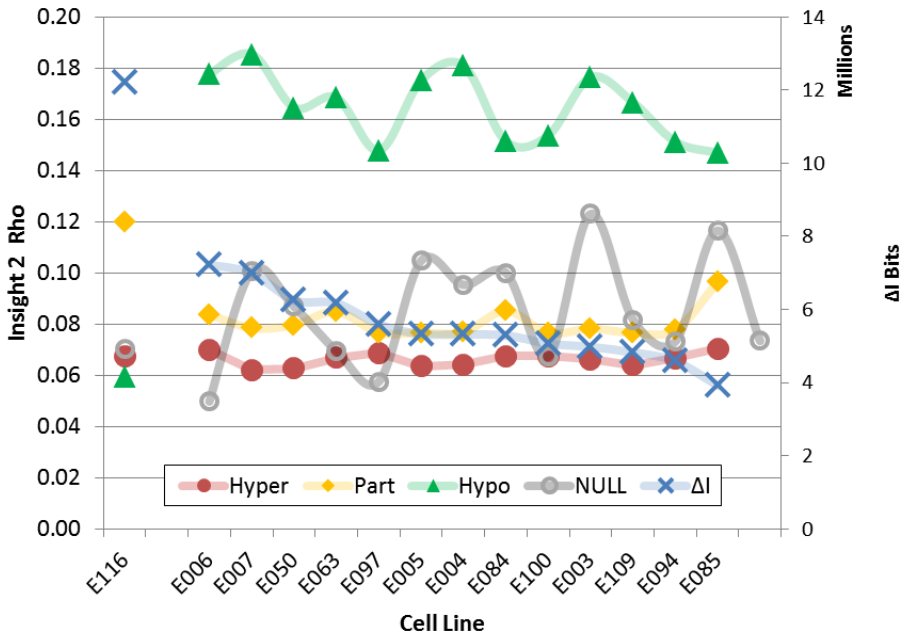


Figure 2.16: Raw WGBS marginal information for 4 covariate classes. Displayed the marginal information (blue points, right vertical axis) provided by genomic segmentation for a variety of cell types (horizonal axis) including E003 (H1 hESC) a common reference cell-type. The four classes consisted of hypomethylated (green), hypermethylated (red), partially methylated (yellow) and unclassified (grey), each of these is described by the value of ρ for the respective class (left vertical axis). This 4-way classification showed little differentiation among the three classes with lower ρ values. The significant variation at the E116 cell-type was due to a flawed data source that was replaced with imputed data. This partitioning scheme was discarded in favor of a simpler hypomethylated/non-hypomethylated bipartition over uniformly imputed data.

Hypomethylated loci were identified using the HMR software from the Methpipe package V3.3.1, downloaded from the Smith laboratory at USC at http://smithlabresearch.org/software/methpipe/ . HMR uses a hidden Markov model to

identify hypomethylated regions from collections of CpG methylation counts, provided by the imputed Roadmap data. HMR software can be used to identify hypomethylated, hypermethylated and partially methylated regions from methylation levels at CpG dinucleotides.

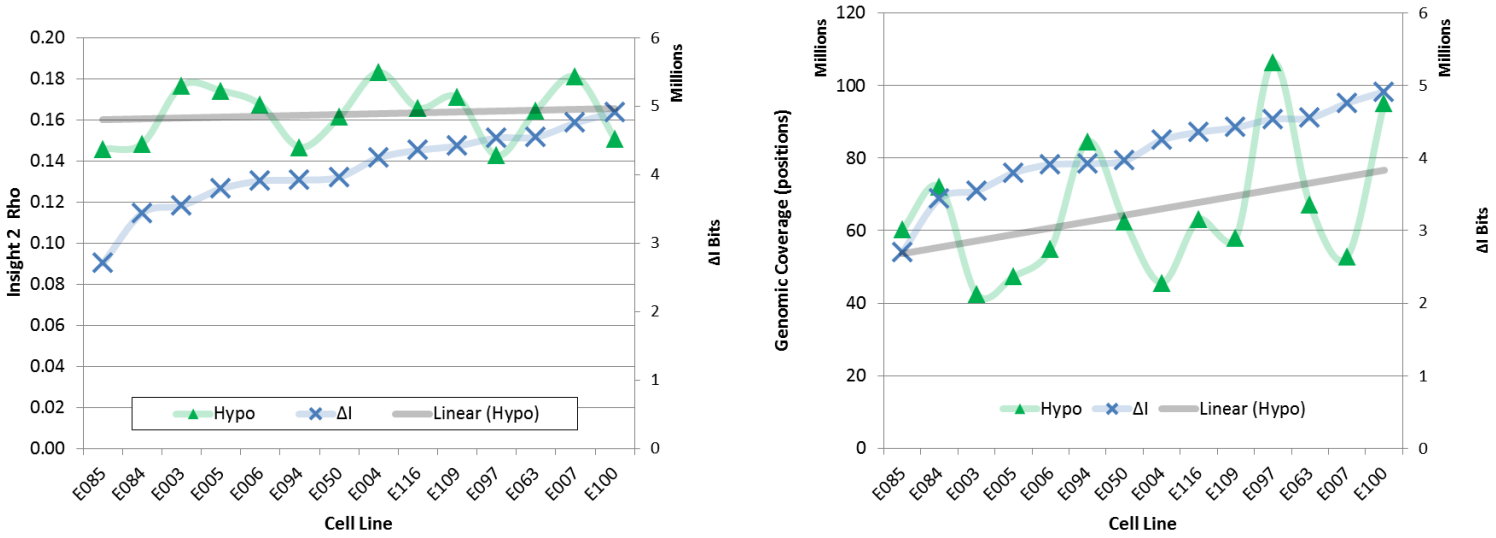


Figure 2.17: Imputed WGBS quantization. Imputed hypomethylation data from Roadmap shows the value of ρ at hypomethylated retions associated with transcriptional activity (left, green data laft vertical axis). Values of ρ vary from 0.14 to 0.18 across cell-types (horizonal axis). At right the genomic coverage of the hypomethylated WGBS class varies from 40-110 million positions of coverage, depending on methylation markers in each cell type. Values for ρ and coverage are more constant than the previously examined data, especially at cell type E116.

2.5.2.6 Chromatin state

Chromatin state data was downloaded from the following location:

http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/E{XXX}_25_imputed12marks_mnemonics.bed.gz ,

where {XXX} is replaced with the 127 valid Roadmap cell-type IDs in the range 001-129. This data is developed as a segmentation of each Roadmap cell type into 25 states, based on a hidden Markov model using 12 imputed chromatin marks as emission parameters. The states, imputed marks, and emission probabilities from are provided blow for convenience[100].

Emission Parameters

	H3K9me3	H3K36me3	H4K20me1	H3K79me2	H3K4me1	H3K27ac	DNase	H3K9ac	H3K4me3	H3K4me2	H2A.Z	H3K27me3
1_TssA	0.4	0.1	0.0	5.0	0.4	89.3	92.3	96.5	99.9	99.1	86.4	3.5
2_PromU	0.6	0.0	0.5	6.2	99.3	91.5	82.4	96.1	99.6	100.0	95.6	23.4
3_PromD1	0.2	1.7	7.0	98.4	60.8	99.8	92.2	100.0	100.0	100.0	93.2	6.3
4_PromD2	0.9	7.3	20.3	94.2	87.5	52.3	8.0	55.1	86.7	98.0	7.2	5.2
5_Tx5'	0.4	1.1	17.0	76.2	0.3	0.2	0.7	0.0	0.0	0.1	0.0	0.1
6_Tx	0.7	94.2	45.7	80.9	7.2	1.0	1.0	0.0	0.0	0.4	0.0	0.1
7_Tx3'	0.1	85.8	1.3	0.8	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0
8_TxWk	0.0	2.4	0.1	1.3	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0
9_TxReg	0.2	27.5	60.2	98.1	98.3	99.9	72.1	92.8	74.0	99.6	6.1	1.3
10_TxEnh5'	0.2	25.9	49.5	96.2	94.1	94.6	25.7	5.8	2.0	41.6	0.5	0.2
11_TxEnh3'	0.4	89.6	14.8	11.0	74.3	50.0	20.3	2.9	1.4	11.8	0.5	0.5
12_TxEnhW	0.1	9.3	48.3	95.5	76.8	3.2	6.6	0.0	0.4	18.9	0.1	0.8
13_EnhA1	0.2	4.0	1.3	5.9	99.3	99.9	83.7	95.7	38.3	95.7	43.3	0.4
14_EnhA2	0.2	0.5	0.8	2.6	97.4	97.2	59.1	7.5	9.6	96.8	29.0	0.6
15_EnhAF	0.3	0.4	0.5	2.1	97.7	94.5	31.0	3.7	1.2	2.3	6.3	0.7
16_EnhW1	0.1	0.0	0.2	0.5	91.2	16.8	39.1	3.4	15.2	73.3	46.4	0.9
17_EnhW2	0.1	0.2	0.5	1.0	75.9	0.4	13.8	0.0	0.0	1.3	0.9	0.5
18_EnhAc	0.3	0.3	0.1	1.1	4.9	64.3	19.4	0.7	0.5	3.3	1.2	0.4
19_DNase	0.1	0.0	0.1	0.0	3.4	0.3	44.7	0.0	0.0	1.4	6.2	0.1
20_ZNF/Rpts	88.9	82.0	1.0	15.9	0.5	0.1	0.6	0.2	4.7	1.4	0.0	0.1
21_Het	69.6	0.2	0.0	0.0	0.1	0.0	1.4	0.0	0.2	0.1	0.0	0.5
22_PromP	2.6	0.3	0.2	2.0	9.6	11.0	19.6	9.1	34.5	67.6	18.0	1.0
23_PromBiv	2.2	0.3	2.4	4.0	76.6	15.6	29.5	23.9	63.9	83.1	44.4	96.6
24_ReprPC	1.1	0.1	0.3	0.3	3.4	0.2	1.2	0.0	0.1	0.3	0.1	72.4
25_Quies	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1

Median Enrichments

	Genome %	CpG.hg19	Exons.Gencodev10.hg19	Genes.Gencodev10.hg19	Introns.Gencodev10.hg19	TES.Gencodev10.hg19	TES_2kbp.Gencodev10.hg19	TSS.Gencodev10.hg19	TSS_2kbp.Gencodev10.hg19	ZNF_genes
1_TssA	0.18	97.56	10.08	1.32	0.64	3.50	2.55	96.91	9.44	3.58
2_PromU	0.41	36.50	4.75	1.12	0.84	2.94	2.37	16.12	7.18	1.86
3_PromD1	0.41	55.24	8.53	1.82	1.30	3.55	2.74	36.57	9.70	4.26
4_PromD2	0.19	2.32	3.44	1.91	1.80	3.16	2.51	2.52	7.98	4.65
5_Tx5'	2.22	0.13	0.56	1.97	2.08	0.72	0.94	0.59	1.22	2.44
6_Tx	0.70	1.34	5.15	1.96	1.72	5.95	3.83	2.52	2.84	3.25
7_Tx3'	3.48	0.95	5.57	1.93	1.65	5.93	4.34	1.99	2.58	2.76
8_TxWk	5.88	0.34	1.87	1.85	1.86	1.70	2.15	0.84	1.55	2.40
9_TxReg	0.30	2.76	4.01	1.93	1.77	4.20	2.60	3.85	5.05	1.45
10_TxEnh5'	0.38	0.36	1.81	1.96	1.97	2.00	1.54	1.48	1.96	1.41
11_TxEnh3'	0.21	1.22	7.20	1.89	1.47	7.26	4.79	2.65	3.04	1.35
12_TxEnhW	0.51	0.28	1.15	1.96	2.03	1.18	1.18	1.05	2.27	2.09
13_EnhA1	0.22	0.93	2.16	1.26	1.18	1.89	1.77	2.76	2.50	0.79
14_EnhA2	0.34	0.37	1.45	1.22	1.20	1.29	1.33	1.84	1.95	0.83
15_EnhAF	0.48	0.16	1.31	1.23	1.25	1.17	1.31	1.06	1.59	0.69
16_EnhW1	0.28	1.78	1.70	0.99	0.94	1.58	1.39	2.99	2.98	1.01
17_EnhW2	0.95	0.25	1.23	1.24	1.25	1.17	1.24	1.02	1.48	0.79
18_EnhAc	0.27	0.24	1.14	1.21	1.22	1.05	1.23	1.33	1.57	0.66
19_DNase	0.63	0.43	0.92	0.94	0.95	1.04	0.96	1.70	1.16	0.52
20_ZNF/Rpts	0.18	0.68	5.02	1.86	1.61	4.02	3.30	1.03	1.50	71.79
21_Het	0.91	0.91	1.02	0.71	0.69	0.86	0.83	0.55	0.76	7.69
22_PromP	0.20	14.16	3.01	1.22	1.09	2.22	1.79	10.48	5.02	1.51
23_PromBiv	0.25	53.53	5.88	1.31	0.95	3.71	2.62	12.96	6.81	0.72
24_ReprPC	1.32	4.88	2.02	0.99	0.92	1.74	1.69	1.69	2.75	0.47
25_Quies	78.38	0.14	0.53	0.83	0.85	0.58	0.67	0.38	0.64	0.50

The conditional relationship between chromatin state and selective pressure is not clear a-priori, so chromatin state is treated as a non-monotonic covariate. For each FitCons2 decision tree subdividion, the expected value of ρ is recalculated for each chromatin state. The chromatin states are then sorted by this expected ρ and the resulting ordering is treated as a monotonic covariate along the sorted chromatin

classes. Only bi-partitions consistent with this ordering are explored, limiting the number of candidate partitions investigated at each node for this covariate to the 25 provided chromatin classes.

2.5.2.7 CDS annotation

CDS annotation is drawn directly from GencodeV19, downloaded from: ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz .

CDS covariates are extracted as per INSIGHT[66]. INSIGHT scores for all positions annotated in each phase, as well as positions annotated as being in a "start" or "stop" codon. As the stop codon selective pressure was not substantially different from the CDS average, only start codon and phasing categories were used. When a position was annotated as being in more than one category, it was assigned to the category with the highest INSIGHT-ρ . CDS positions 1, 2 and 3 descried in this paper correspond to CDS "phases" 0, 1, and 2 respectively.

CDS Class	INSIGHT-ρ
cds-start	0.8125
cds-ph-1	0.7015
cds-ph-0	0.6499
cds-stop	0.6460
cds-ph-2	0.5712

2.5.2.8 Splicing annotation

Splice sites were inferred from each CDS/Exon boundary observed in the CDS dataset from GENCODE V19 (see, 2.5.2.7 CDS annotation). Conservation is not a monotonic property of distance from splice site. To capture this property as a covariate, intronic positions at each fixed distance within 50 bp of a CDS splice site

were combined and INSIGHT-ρ was generated for the collection of positions at each integral distance. Distances from 5’ and 3’ splice sites were treated separately for this calculation. Distances from splice sites were then sorted by ρ value and a maximum likelihood quantization of ρ values was determined for 4 classes. As only the 2 exonic positions upstream from the 5’ splice site seemed under elevated selective pressure, only those exonic positions were added to the splice covariate.

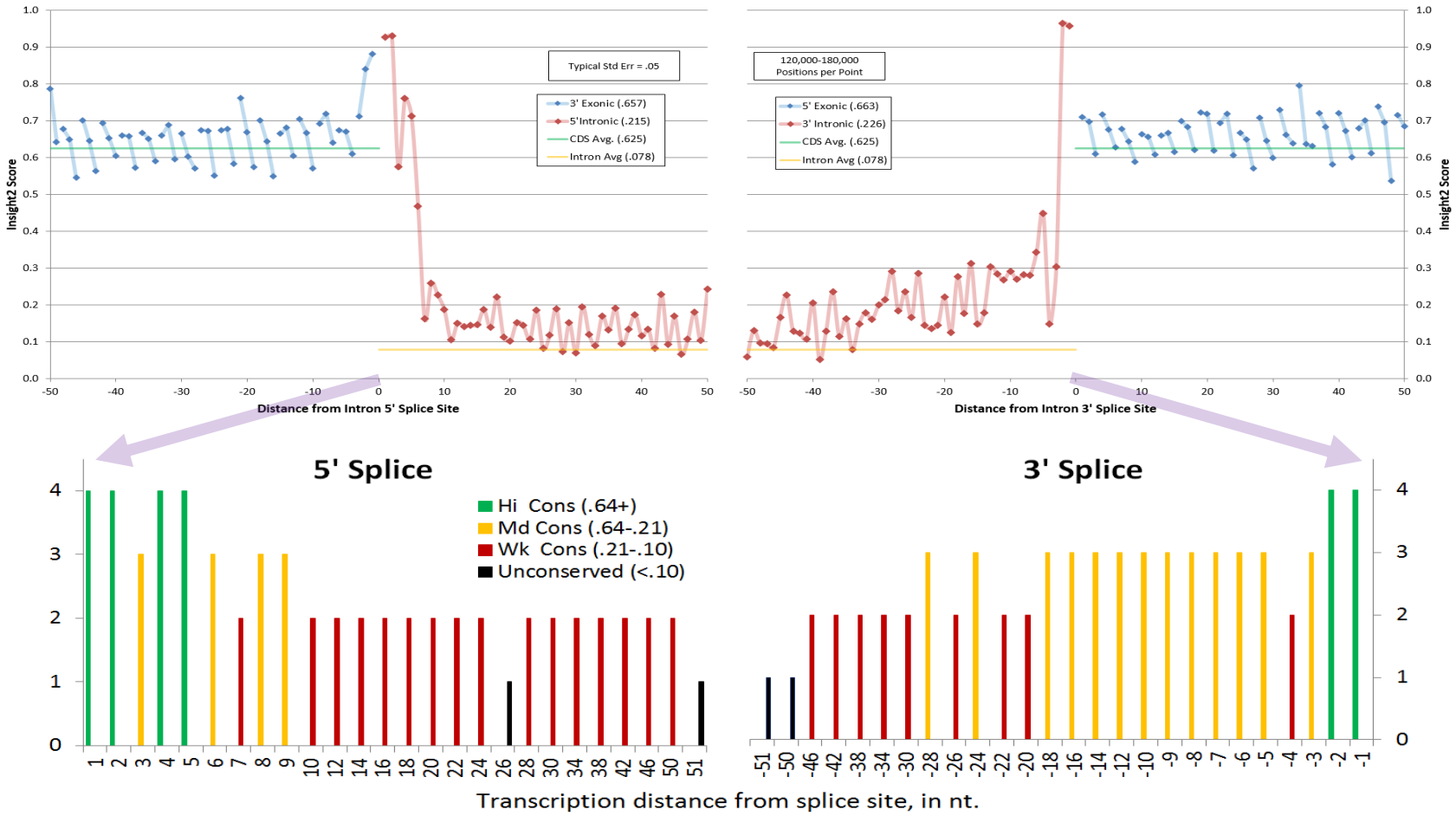


Figure 2.18: Intronic splicing covariate quantization. Before (above) and after (below) quantization. The splicing covariate identifies selective pressure at various distances from 5' and 3' splice sites. Above, Y axis values are INSIGHT scores, and both exonic and intronic positions are shown. For intronic positions (below) Y axis values are quantized class numbers, while range of INSIGHT-ρ scores covered by each class appears in legend.

2.5.2.9 Melting temperature annotation

DNA melting temperatures were downloaded from:

http://meltmap.uio.no/rawdata-hg18/chr{X}.dat.bz2 ,

where {X} is replaced with chromosome numbers 1-22, with one file for each autosomal chromosome. An overview page may be found at

http://meltmap.uio.no/rawdata-hg18.html . Note this data set is in hg18 coordinates and the `liftOver` tool[101] was used to convert it to hg19.

DNA melting temperature provides insight into of the amount of energy necessary to separate the strands of DNA. Lower melting temperature indicates that the primary sequence may be more accessible to biochemical activity, while higher melting temperature may indicate greater chemical stability. Melting temperature in a window of contiguous positions is strongly correlated with the fraction of guanine and cytosine nucleotides, as opposed to adenine or thymine. This fraction is generally referred to as "GC content". Local DNA melting temperature in degrees centigrade was obtained from the Human Genomic Melt Map for hg18 and mapped to hg19 using the liftOver tool. Temperature was quantized as per other continuous covariates, and it was found that both extremely high and extremely low melting temperatures had elevated selective pressure. This annotation was quantized into 5 categories to reflect the extreme and intermediate values: VeryLow, Low, Medium High and VeryHigh.

Table 2.2: DNA melting temperature. DNA melting temperature annotation, FitCons2 covariate class, INSIGHT-ρ for each class, and corresponding standard error in ρ estimate.

Class	Desc	Range °C	# positions	ρ	Δρ
1	Very Low	-61.5)	21,926,440	0.3375	0.0139
2	Low	[61.5-63.5)	90,001,320	0.1647	0.0037
3	Typical	[63.5-76.5)	2,657,043,760	0.0706	0.0005
4	High	[76.5-80.5)	94,187,385	0.1556	0.0038
5	Very High	[80.5-	17,874,381	0.3436	0.0133

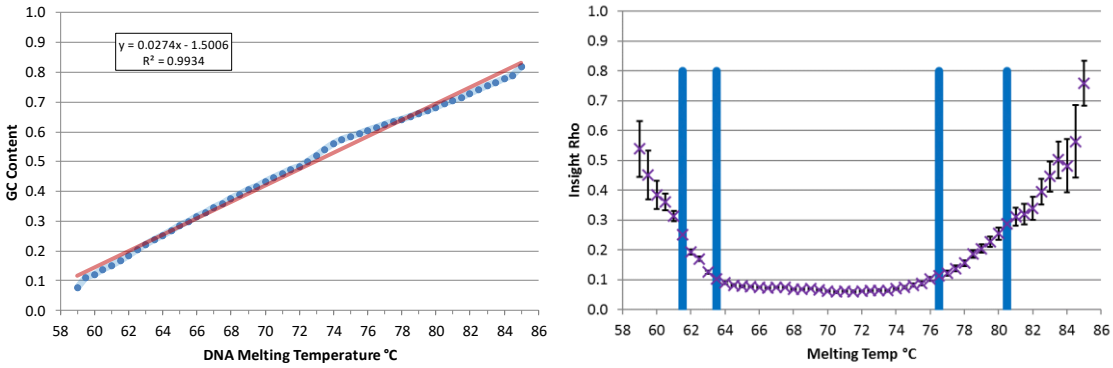


Figure 2.19: Melting covariate properties and quantization. Fraction of positions at each melting temperature that are G or C nucleotides (left). Estimated INSIGHT-ρ for positions at each melting temperature, binned by 0.5°C increments (right). Blue lines at right show covariate bin boundaries determined by maximum likelihood.

2.5.2.10 Transcription factor binding site pseudoannotation

Transcription Factor Binding Sites (*TFBS*) were obtained from two sources, Leo Arbiza[28] and the Ensembl Regulatory build V84[60].

The larger more informative source was obtained from Leo Arbiza, and may be viewed via the CSHL mirror of the UCSC genome browser. The data source for this set is presently:

Nextgen:/data/projects/fncSel/1_regions/tfbs/outdated/noReplicates_islands_filtered/hg19_trimmed_tfbs_beds/ .

There are data for 86 TFs, identifying 2,731,535 binding sites that span a total of 19,619,025 unique autosomal positions, in 23 cell types consisting of: GM12878, GM12891, GM12892, H1-HESC, HCT-116, HEK293, HEK293-T-REX, HELA-S3, HEPG2, HUVEC, K562, MCF-7, NB4, NT2-D1, PANC-1, PBDE, PBDEFETAL, PFSK-1, SH-SY5Y, SK-N-MC, SK-N-SH, SK-N-SH_RA, and U87. Many of these cell types are not Roadmap cell types.

List of 86 Transcription Factors (*TF*s) from Arbiza Set

AP-2ALPHA	C-JUN	FOXA2	MAX	RAD21	TR4
AP-2GAMMA	C-MYC	FOXP2	MEF2A	RFX5	USF1
ATF3	CTCF	GABP	MEF2C	RPC155	USF2
BAF155	CTCFL	GATA1	NANOG	SIX5	WHIP
BAF170	E2F1	GATA2	NF-E2	SMC3	YY1
BATF	E2F4	GATA3	NF-YA	SP1	ZBTB33
BCL11A	E2F6	GCN5	NF-YB	SP2	ZBTB7A
BCL3	EBF1	GTF2F1	NRF1	SRF	ZEB1
BCLAF1	EGR1	HNF4A	NRSF	STAT1	ZNF143
BDP1	ELF1	HNF4G	P300	SUZ12	ZNF263
BHLHE40	ELK4	IRF3	PBX3	TAL1	ZNF274
BRCA1	ETS1	IRF4	POU2F2	TCF12	
CEBPB	FOSL1	JUND	POU5F1	TCF7L2	
C-FOS	FOSL2	MAFF	PRDM1	TFIIIC-110	
CHD2	FOXA1	MAFK	PU1	THAP1	

Of these, 8 were removed as being associated with non-sequence-specific binding: ZBTB33, THAP1, BDP1, GCN5, E2F1, RPC155, SMC3, P300. After removal, this left 2,595,018 binding sites covering 19,245,632 unique positions. The PWM's for each TF was obtained from personal communication with Arbiza 3-Jun-2016 (pwms.txt) and are stored in nextgen:/data/projects/fitcons2/src/tfbs/leo2/src/motif/ in the file pwms-raw.txt.

The PWM's were converted to an information score in the range 0.0-2.0 bits, for each oriented motif position in each identified TFBS. Positions mapping to more than one TFBS used the highest information value. This generated single value for each genomic position, with a value of 0 being used for positions outside TFBS. This real-valued score was quantized into 4 classes using the General Quantization Method listed above, providing a marginal information content of 5,100,167 bits.

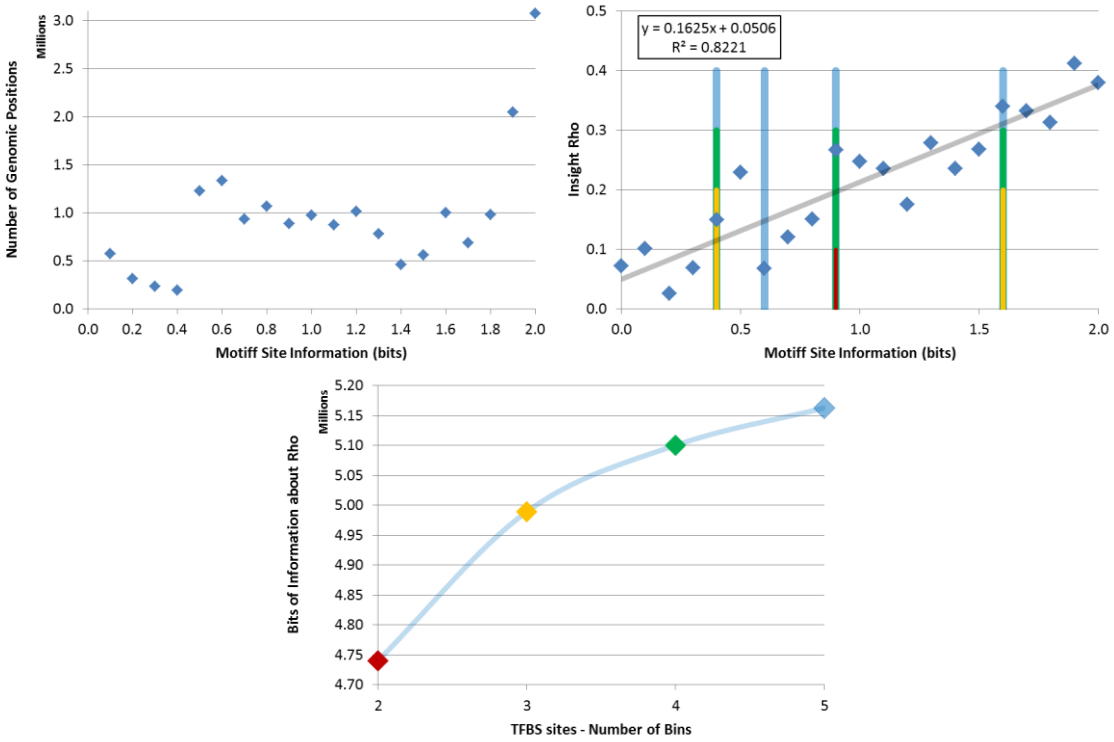


Figure 2.20: Arbiza binding site distributions. Motif information at each genomic position in a TFBS ranges from 0.0-2.0 bits in increments of 0.1 bits along the horizonal axis of the two upper sub figures. An information value of 0.0 indicates an equal distribution of the 4 nucleotides across instances of that motif, a value of 2.0 indicates that the same nucleotide is always found at the corresponding motif position. The upper left, panel shows the number of genomic positions in Millions of bp (vertical axis) for each level of motif information. The upper right panel shows the correlation of INSIGHT-ρ on the vertical axis with motif information. Also shown are the partitions of motif information that are most informative about ρ for: 2 bins (red), 3 bins (yellow), 4 bins (green) and 5 bins (blue). The bottom panel shows the increasing information about ρ provided by increasing the number of bins. While five bins provide more information about ρ than four, four bins were selected as a balance between information and intelligibility.

The Ensembl data set is available from:

ftp://ftp.ensembl.org/pub/grch37/release-84/regulation/homo_sapiens/MotifFeatures.gff.gz

This data set consisted of binding sites for 48 TFs at 604,066 loci covering 4,467,614 unique autosomal positions.

List of 48 Transcription Factors (*TF*s) from Ensembl Set

BHLHE40	CJUN	CTCF	E2F1	E2F6	EGR1
CFOS	CMYC	CTCFL	E2F4	EBF1	ELF1

ETS1	GATA1	JUND	NRSF	SP2	TR4
FOSL1	GATA2	MAX	PAX5	SREBP1	USF1
FOSL2	HNF4A	MEF2A	PBX3	SREBP2	YY1
FOXA1	HNF4G	MEF2C	POU2F2	SRF	ZBTB33
FOXA2	IRF4	NFKB	PU1	TCF12	ZEB1
GABP	JUNB	NRF1	SP1	THAP1	ZNF263

Of these, 4 were removed as being associated with nonspecific binding: ZBTB33, THAP1, E2F1. This left 588,958 binding loci covering 4,363,019 unique autosomal positions. The list of cell types from which binding data was generated was not provided. For each TFBS a JASPAR[102] PWM ID similar to the form "MA0366.1" was provided. JASPAR PWMs were downloaded from http://jaspar.genereg.net/html/DOWNLOAD/all_data/FlatFileDir.tar.gz. Component matrices carried a file date of 04-Dec-2015 and are presumably from the 2016 release of the JASPAR database. Positions in TFBS were scores as per the Arbiza data set, and a quantization into 4 classes was performed, providing a marginal information content of 2,037,030 bits.

The two sets of TFBS classes (Arbiza & Ensembl) were then combined into 4 joint classes using the general aggregation method described above (2.5.2.2 General aggregation method), producing the 4-class TBFS pseudoannotation, providing a marginal information content of 5,289,220 bits, a value higher than either data set alone.

2.5.2.11 Small RNA-seq pseudoannotation

Short RNA-seq provides access to transcription activity that may detect biological micro RNAs that are not typically identified by standard RNA-seq. Only some of the Roadmap cell lines had small RNA-seq data, so available sources were combined into a single pseudoannotation. Data from 5 cell-sources was selected corresponding to CD20, HUVEC, PKF, BGM and hESC (4Star). Multiple small replicates and extraction chemistries were available, so representative samples were blended using the general quantization method (2.5.2.1 General quantization method)

on read depth, followed by the general aggregation method (2.5.2.2 General aggregation method). The order of processing was:

- Replicates were quantized individually
- Replicates were aggregated within a chemistry & cell source
- Chemistries were aggregated within a cell source
- BGM & hESC were aggregated into a first group
- PFK, HUVEC and CD20 were aggregated into a second group
- Groups one and two were aggregated into the covariate.

The data and sources for the small RNA-seq data are as follow:

ENCODE CD20

CD20 consists of B lymphocyte blood cells responsible for producing antibodies, from two donors.

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlShortRnaSeq

Files:
wgEncodeCshlShortRnaSeqCd20ro01794CellShorttotalTapMinusRep1.bigWig
wgEncodeCshlShortRnaSeqCd20ro01778CellShorttotalTapMinusRep2.bigWig
wgEncodeCshlShortRnaSeqCd20ro01794CellShorttotalTapPlusRep1.bigWig
wgEncodeCshlShortRnaSeqCd20ro01778CellShorttotalTapPlusRep2.bigWig

ENCODE HUVEC

HUVEC is a human umbilical vein endothelial cell line.

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlShortRnaSeq

Files:
wgEncodeCshlShortRnaSeqHuvecNucleusShorttotalTapMinusRawRep3.bigWig
wgEncodeCshlShortRnaSeqHuvecNucleusShorttotalTapMinusRawRep4.bigWig
wgEncodeCshlShortRnaSeqHuvecNucleusShorttotalTapPlusRawRep3.bigWig
wgEncodeCshlShortRnaSeqHuvecNucleusShorttotalTapPlusRawRep4.bigWig

UCSF PKF (From EDACC / Baylor College of Medicine)

Penis Foreskin Keratinocytes (PKF), an epidermal (skin) cell that produces keratin.

https://www.genboree.org/EdaccData/Release-9/experiment-sample/smRNA-Seq/Penis_Foreskin_Keratinocyte_Primary_Cells/

UCSF-UBC.Penis_Foreskin_Keratinocyte_Primary_Cells.smRNA-Seq.skin01.bed.gz
UCSF-UBC.Penis_Foreskin_Keratinocyte_Primary_Cells.smRNA-Seq.skin02.bed.gz
UCSF-UBC.Penis_Foreskin_Keratinocyte_Primary_Cells.smRNA-Seq.skin03.bed.gz

UCSF BGM (From EDACC / Baylor College of Medicine)

Brain Germinal Matrix (BGM) is the source of neurons & glial cells most active ar 8 28 weeks gestation.

https://www.genboree.org/EdaccData/Release-9/experiment-sample/smRNA-Seq/Brain_Germinal_Matrix/

UCSF-UBC.Brain_Germinal_Matrix.smRNA-Seq.HuFGM01.bed.gz
UCSF-UBC.Brain_Germinal_Matrix.smRNA-Seq.HuFGM02.m09072.bed.gz

UCSF 4Star

UCSF 4Star (4*) is an embryonic stem cell line (A11831)

https://www.genboree.org/EdaccData/Release-9/experiment-sample/smRNA-Seq/UCSF-4star/

UCSF-UBC.UCSF-4star.smRNA-Seq.m05795.bed.gz
UCSF-UBC.UCSF-4star.smRNA-Seq.m05796.bed.gz

2.5.3 Tree complexity and refinement

The FitCons2 covariate decomposition tree was calculated to a minimum cutoff of 5 bits. Recursive refinement of the tree was halted when no covariate

bipartition would yield a conditional improvement in negative log likelihood (*NLL*) of at least 5 bits (3.47 nats). Tree decomposition produces a submodel with each division, adding 1 parameter for the bipartition and 3 parameters for an additional INSIGHT model. The decomposition tree was pruned to 50 bits, so any node that could not provide an improvement in conditional likelihood of at least 50 bits under bipartitioning, was taken as terminal (a leaf of the tree). This would correspond to a likelihood ratio test statistic (D) of 69.4, which would provide a p$< 3.1 \cdot 10^{-14}$ over a null model for submodel with 4 additional degrees of freedom and a $p < 0.01$ for even 40 new parameters. Pruning the FitCons2 tree to 50 bits also improved intelligibility of the leaves by limiting their number to 61 classes. A cutoff of 5 bits allowed for 195 leaves/classes. The relationship between cutoff, number of leaves and average fraction of sites under selection is provided in the figure below.

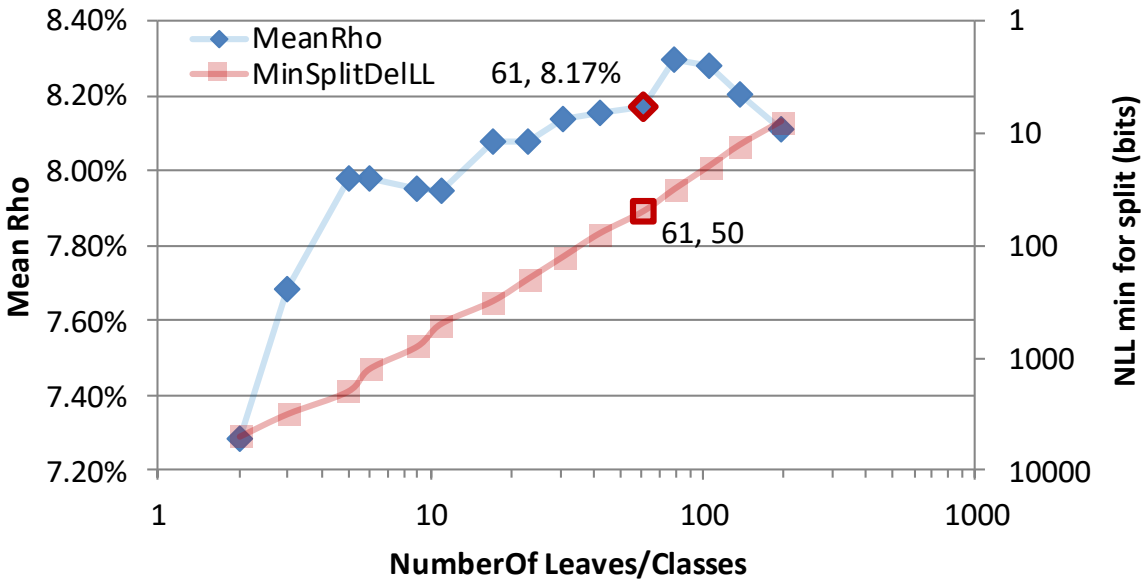


Figure 2.21: Relationship between pruning, leaves and average ρ. The horizonal axis (log scale) represents the number of leaves, or terminal classes, in the FitCons2 tree decomposition as the termination threshold is decreased logarithmically from 5,000 bits to 5 bits (right vertical axis, log scale). As the threshold decreases, the number of leaves increases (red curve). Similarly, the average value of ρ measured across leaves rises steadily as the cutoff declines from 5,000 bits to 32 bits, then ρ declines slightly (blue curve). The cutoff utilized in this work (50 bits) blends conservative significance testing with model simplicity, and is represented by the data points outlined in red.

2.5.4 Browser display of scores, classes and covariate data

The primary tool for viewing covariates, classes and scores is the UCC genome browser and its mirrors. Data is provided in structure called a track-hub which can be accessed by any web browser and displayed by the genome browser's display logic. The genome browser arranges individual data elements into rows with one column per genomic position. Examples of browser displays are provided in Figure 2.12 and Figure 2.13. This browser track hub may be accessed by researchers via:

http://genome.ucsc.edu/cgi-bin/hgHubConnect?hgHub_do_redirect=on&hgHubConnect.remakeTrackHub=on&hgHub_do_firstDb=1&hubUrl=http://compgen.cshl.edu/fitCons2/hub.txt

Select the "go" option when the redirection completes. By default, one example cell-type for each of the 9 default tissue classes is provided, along with the cell-type integrated scoring. An alternative repository for FitCons related research can be found via the author's web site www.fitcons.science [84].

2.5.5 Information theoretic properties of covariates

To assess the informative properties of individual covariates in the FitCons model, we performed four sets of measurements. The first was an estimate of the information provided by each covariate in the decision tree learned by the standard training, this estimated the amount of information provided by each property in the actual model. We then retrained the model on each single variable, providing an upper bound on the information provided by each covariate. To estimate the amount of unique information provided by each property, we retrained the model holding out each variable, allowing other related genomic properties to be selected in lieu of the held-out variable. Finally, we approximated an experimental setting in which CDS, RNA-seq and DNase-seq were known and retrained the model several holding out each of the remaining properties, individually.

While information is measured as a change in base two negative log likelihood genome wide (*NLL*, in bits), covariates tend to separate the genome into a large collection of positions with weak selective pressure (e.g. null) and a smaller set of positions with one (WGBS) or more classes of more strongly selected positions. To provide a sense of the information density of each property, we also calculated a pseudo-entropy, by dividing this change in log-likelihood by the number of observations in the non-null class. The NLL and entropy measures indicated herein are based on INSIGHT-NLL, not the Shannon information about S used to develop covariate quantization ($S \in \{0,1\}$ where $S = 1$ means a site is under selection and at each genomic position, $\rho = [S]$ over a class of N positions). While these information measures are strongly related, they are represented on different scales and cannot be directly compared. Individually, CDS annotation was the single most informative covariate both in terms of total information (>35 Kbits) and information density (>1000 μbits/bp). This property is unsurprising as CDS are both highly conserved relative to non-coding regions and are compact, covering only ~1.3% of autosomal positions. The next most informative covariates were RNA-seq and (~20 Kbits) followed by chromatin state (~10 Kbits). However, the most densely informative covariates were splicing (>200 μbits/bp) and transcription factor binding (~100 μbits/bp). While less informative genome wide than RNA-seq and chromatin state, the small number of sites involved in splicing and TF binding made them an order of magnitude more densely informative than RNA-seq and chromatin state.

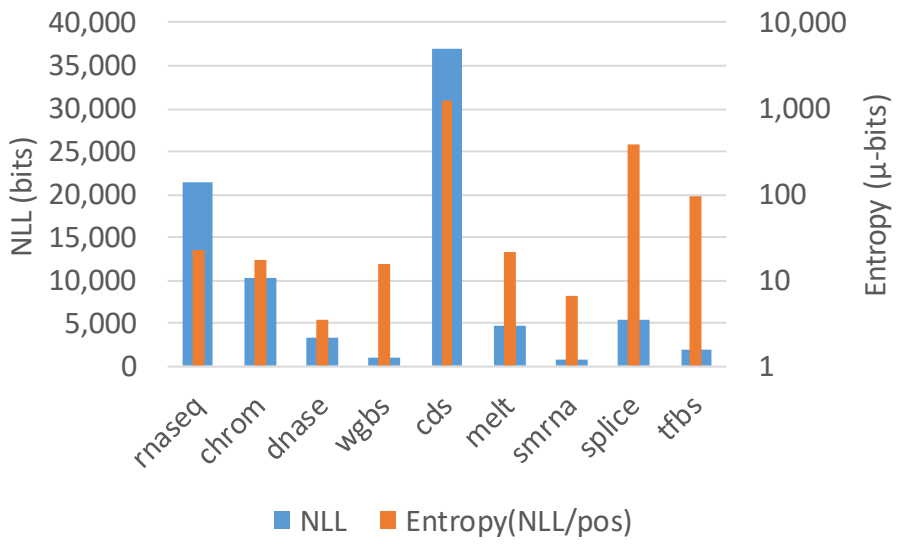


Figure 2.22: Covarite information in FitCons2 Information in INSIGHT-NLL per covariate (horizonal axis) used in the FitCons2 tree decomposition. Blue bars represent total information per covariate (left vertical axis) while orange bars represent an entropy that attempts to capture information density of covariate, measured on logarithmic axis on right.

The amount of unique information per covariate may prove even more interesting as it represents information about selective pressure that is not recoverable from combinations of other properties. CDS annotation again had the highest value. While RNA-seq recovered much of the information provided by CDS, the clear boundaries and phasing structure added >10 K bits of unique information. The next most irreplaceable covariate was melting temperature, corresponding to broad highly conserved regions at with both extremely high and extremely low GC dinucleotide content. Following melting temperature were RNA-seq and chromatin state, each with about 2000 bits of cell-type specific information that could not be recovered from other sources.

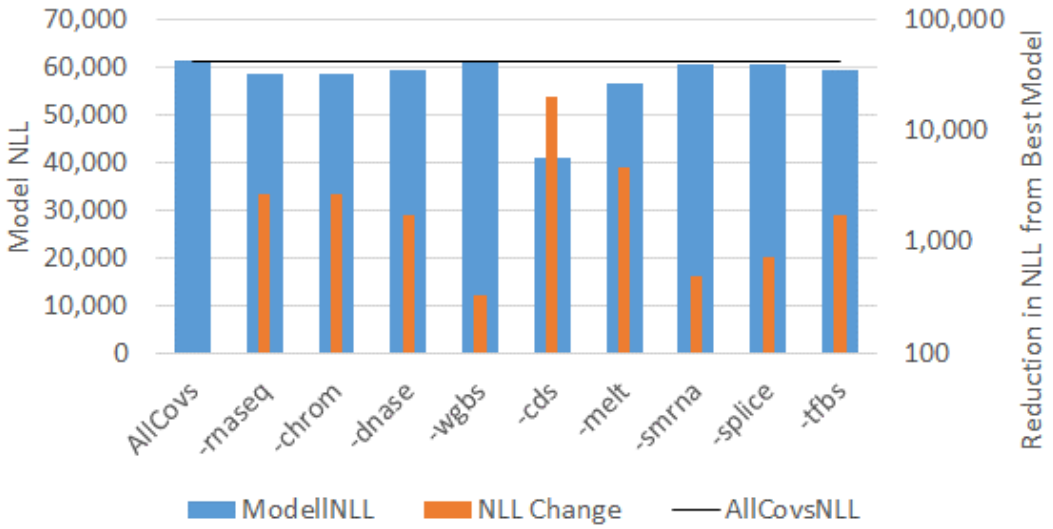


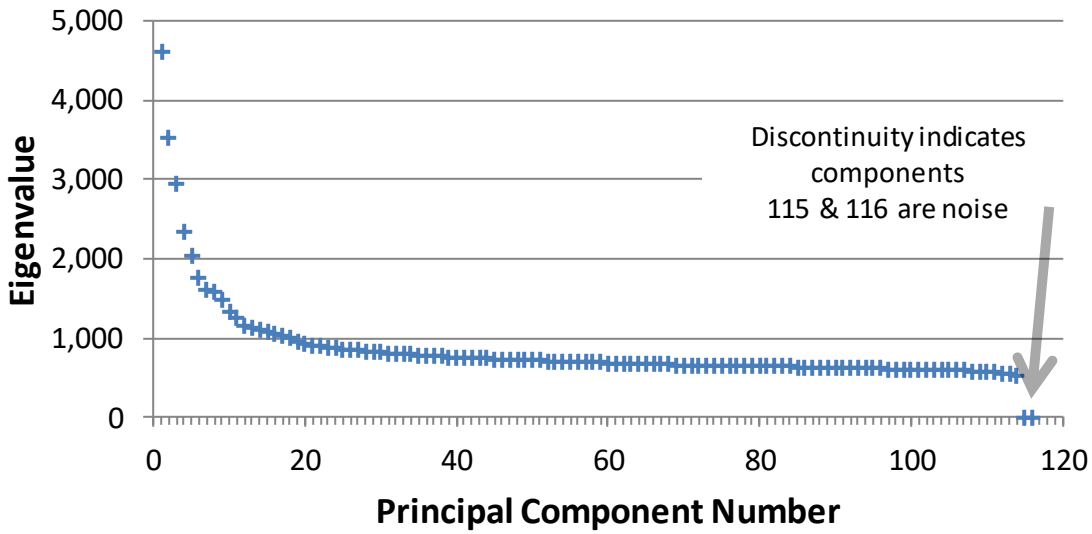
Figure 2.23: Unique information per covariate in FitCons2. Blue bar at left represents improvement in NLL for complete FitCons2 model (left vertical axis). Successive bars (horizonal axis) represent the model retrained with a single covariate missing. Removal of any covariate reduces NLL of trained model (blue bars). However, some covariates like CHROM have information that can readily be derived from other covariates resulting in a modest NLL impact. Other covariates, like CDS contain information that is not retrieved from remaining covariates resulting in greater impact on trained NLL. For contrast, orange bars measure the distance between the retrained and complete covariate NLL (horizonal black line at top) on the logarithmic right vertical axis.

2.5.6 Cell type independent scoring

2.5.6.1 Cell-type information weighting

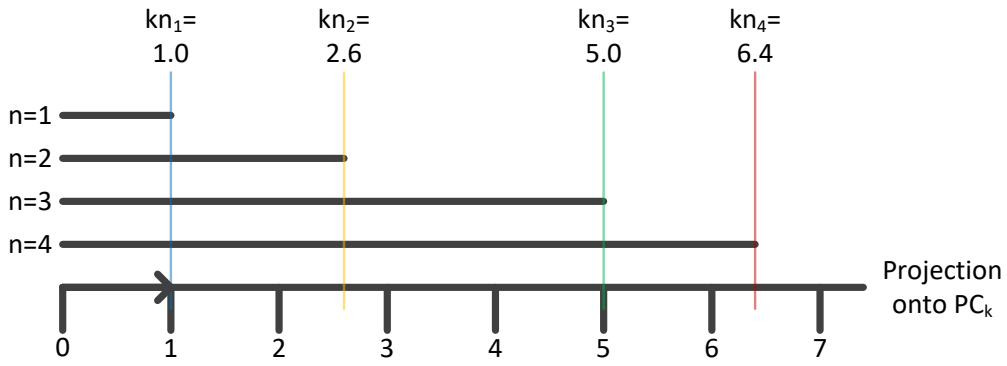
A principal components analysis (PCA) of scores at each genomic position was performed. This analysis utilizes the 2,881,033,286 autosomal positions in hg19 as distinct variables. Each of the 115 cell-types was treated as a separate observation of the 2.88 billion variables. A naïve PCA decomposition with 3 billion variables is problematic on modern hardware, even with iterative methods. To simplify computation, the dimensionality of this PCA was reduced by observing that there were only 61 unique values among all variable observations. Aggregation by unique combinations of the 115 observations of 61 values identified only 88,230,965 unique

exchangeable combinations from the $> 10^{47}$ possibilities. By weighting each unique combination of observations by the square root of the number of times it was observed, PCA was reduced to a simpler calculation over the exchangeably unique combinations of variable values, or <100Million effective variables. While large, this method is amenable to iterative solution using a modified version of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm implemented in the Chemometrics package for R. For verification, 116 principle components were extracted, of which the last two had very low corresponding eigenvalues indicating they were numerical noise, and the highest 114 components were taken as reasonable approximations. This took 12 days to complete using up to 16 cores and 600GB of memory on a contemporary high memory server.



Once the projections of each cell type onto each of the 114 significant principal components (PCs) is calculated, a weight factor is derived for each cell-type, on each of the K PCs, as follows:

- For a specific $PC_{k\in K}$, calculate the projection of the $\rho_{i,n}$ values for all $n \in N$ cell types onto PC_k . This projection is calculated as a dot product with the unit length PC_k. This intrinsically sums across all genomic positions i, to generate a single real value for each n, namely $proj_{k,n}$.
- Rank the N projections, and treat positive and negative values separately (0 values may be ignored). The maximal extent of cell n's projection onto PC_k, is referred to as a "knot" $kn_{n,k}$ (the k subscript is dropped where clear).



- For each successive knot t, the distance $kn_t - kn_{t-1}$ is evenly distributed as a weight, to all cell types that reach kn_t (and $kn_0 \triangleq 0$). Thus, in this example

  An initial weight of $\frac{kn_1-kn_0}{4} = \frac{1.0-0.0}{4} = 0.25$ is attributed to cell types 1-4.

  An additional weight of $\frac{kn_2-kn_1}{3} = \frac{2.6-1.0}{3} = 0.53$ is added to cell types 2-4.

  An additional weight of $\frac{kn_3-kn_2}{2} = \frac{5.0-2.6}{2} = 0.70$ is added to cell types 3-4.

  An additional weight of $\frac{kn_4-kn_3}{1} = \frac{6.4-5.0}{1} = 1.40$ is added to cell type 4.

  Negative projections are handled separately. Absolute value of distances are used, so all weights are positive.

- After all of the PC are processed this way, we obtain a weight matrix $w_{n,k}$ over the N cell types and K principal components.
- The weight for each cell type w_n is taken as the quadrature sum of all corresponding weights across components, that is

$$w_n = \sqrt{\sum_{k \in K} w_{n,k}^2}$$

- This weight is a measure of the redundancy-compensated magnitude of each centered vector, $\overrightarrow{\rho_n} \triangleq \rho_{\cdot,n}$

If the cell-types had truly orthogonal distributions of $\rho_{i,n}$ across genomic positions i, the total variation could be estimated as ${s_\rho}^2 = \sum_n |\overrightarrow{\rho_n}|^2$. The weight $w_{n,k}$ represents a measure analogous to projection $\overrightarrow{\rho_n} \cdot PC_k$, but with redundant extents counted only once. We can therefore calculate analogous values for $\overrightarrow{w_n} \triangleq w_{\cdot,n}$, $|\overrightarrow{w_n}| = \sqrt{\sum_{k \in K} (w_{n,k})^2}$, and ${s_w}^2 = \sum_n |\overrightarrow{w_n}|^2$.

The ratio $\frac{s_w}{s_\rho}$ is in the range 0.0-1.0 and serves as an estimate of the fraction of non-redundant information in the observations. We then estimate the number of effectively independent observations in the data is $|N| \frac{s_w}{s_\rho}$. In comparing cell types, weights are undefined up to a multiplicative constant. By taking $\bar{\rho} = mean_n(|\overrightarrow{\rho_n}|)$ as a measure of the typical extent of an observation, we can estimate the effective fraction of an independent observation provided by a cell type n as $\frac{|w_{\cdot,n}|}{\bar{\rho}}$, this ratio is referred to as the "relative weight" and has support over the range [0,1]

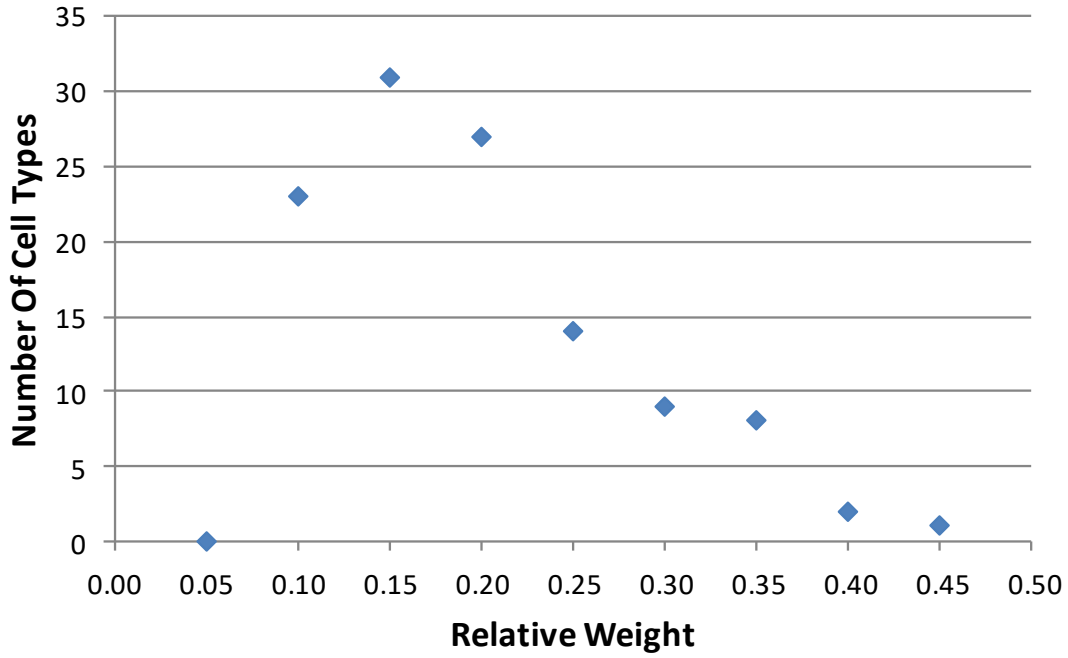


Figure 2.24: Distribution of cell-type weights. Histogram showing the count of cell types (vertical axis) according to cell-type specific weight (horizonal axis) in increments of 0.05. Relative weight is normalized to represent the estimated fraction of a cell type with an independent distribution of ρ along the genome. The sum of all weights is 19.98, suggesting that the actual data set of 115 curated cell-types might be considered equivalent to about 20 independent cell types.

2.5.6.2 Decision tree inference

To insure representation in at both very high and very low values of rho, positions were quantized first into 4 coarse ranges of ρ values (by included lower bound: 0.0, 0.086965, 0.249235, 0.514235). This is the maximum likelihood partition of all scores across-cell types into 4 bins, with number of observations weighted by cell-type specific weighting developed above. The classes with the 3 highest cutoffs were further refined by maximum likelihood subdivision into 4 subclasses generating a total of 1+3x4=13 ordered thresholds.

	Coarse Classes			
Refinements	1	2	3	4
1	01: 0.004727	09: 0.086965	28: 0.249235	46: 0.514235
2		12: 0.124026	33: 0.306676	49: 0.618695
3		18: 0.143616	36: 0.353753	55: 0.707253
4		23: 0.176572	38: 0.403199	57: 0.867059

As described in (Section 2.4.7 Cell-type independent score generation) each of these 13 thresholds was developed into a separate covariate ($C_x(i), x \in \{1, \dots, 13\}$), having a real value for each position in the human genome (i). The value of $C_x(i)$ was calculated as the sum of the weights of all cell types with scores above the threshold, at each position. This process aggregates weights across cell-types, for each of the new covariates C_x.

Each of these 13 covariates was separately quantized into 5 values, and a second FitCons2 run made over this new set of quantized covariates. Using a 50 bit minimum split criterion (the same threshold as the cell-type specific FitCons2 tree), this generated the following cell-type integrated classes (Table 2.3: Cell-type integrated scores).

Table 2.3: Cell-type integrated scores. A listing of the 37 cell-type integrated FitCons2 scores generated by aggregating cell-type specific scores and weights across the 61 FitCons2 classes and 115 Roadmap cell-types. Each score is the ρ parameter from an INSIGHT calculation, and therefore is interpretable directly as expected probability of being under selective pressure. There are a total of 2,881,033,286 genomic positions represented, all of the autosomal hg19 reference. Of these, 1,268,641,360 positions are in the lowest scoring class with a score of 0.038. The expected score is 0.0795, slightly lower than the unweighted expectation of all cell-type specific scores over the same 115 Roadmap cell types. The highest scoring class has a score of 0.884, which is higher than the highest threshold for any cell-type integrated covariate.

ID	INSIGHT-ρ	Genomic Positions	ID	INSIGHT-ρ	Genomic Positions
00	0.884411	1,254,404	18	0.170982	12,716,024
01	0.774583	4,088,374	19	0.169616	27,931,418
02	0.709209	6,725,247	20	0.167544	32,534,569
03	0.641673	10,580,400	21	0.164339	8,824,671
04	0.573412	4,575,827	22	0.161464	4,197,049
05	0.484017	4,154,617	23	0.158777	86,779,593
06	0.405032	1,138,466	24	0.134608	136,496,377
07	0.402870	5,297,598	25	0.117316	15,863,360
08	0.377265	3,562,839	26	0.116710	73,414,419
09	0.349758	1,408,570	27	0.115940	131,590,902
10	0.329534	23,237,399	28	0.104859	74,246,225
11	0.326901	6,919,893	29	0.083666	212,537,714
12	0.243745	14,411,666	30	0.070977	73,352,873
13	0.217752	7,924,061	31	0.067171	150,398,236
14	0.210433	5,707,816	32	0.057540	96,667,693
15	0.181967	5,854,667	33	0.053189	43,417,903
16	0.179468	8,586,928	34	0.050979	170,550,852
17	0.172671	10,205,867	35	0.039302	135,237,409
			36	0.038108	1,268,641,360

2.5.7 Other data sources

Transcription factor binding sites from Ensembl release 84

ftp://ftp.ensembl.org/pub/grch37/release-84/regulation/homo_sapiens/MotifFeatures.gff.gz

HGMD public version HGMD-PUBLIC_20164 from EnsemblV89

http://ftp.ensemblorg.ebi.ac.uk/pub/release-89/variation/gvf/homo_sapiens/Homo_sapiens_phenotype_associated.gvf.gz
NOTE: this is mapped to GRCh38 coordinates, must be converted to CRCh37 = hg19 via liftOver utility.

ClinVar database from dbSNP version 150

ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar_20170501.vcf.gz

now available at

ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive/2017/clinvar_20170501.vcf.gz

REFERENCES

1. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* **470,** 198–203 (2011).
2. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nat. Methods* **5,** 19–21 (2008).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
4. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
5. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).
6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
7. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12,** 628–640 (2011).
8. Mayor, C. *et al.* VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047 (2000).
9. Margulies, E. H., Blanchette, M., Program, N. C. S., Haussler, D. & Green, E. D. Identification and Characterization of Multi-Species Conserved Sequences. *Genome Res.* **13,** 2507–2518 (2003).
10. Boffelli, D. *et al.* Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. *Science* **299,** 1391–1394 (2003).
11. Ovcharenko, I., Boffelli, D. & Loots, G. G. eShadow: A Tool for Comparing Closely Related Sequences. *Genome Res.* **14,** 1191–1198 (2004).
12. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
13. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
14. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of Sequence Conservation at Nucleotide Resolution. *PLOS Comput. Biol.* **3,** e254 (2007).
15. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
16. Graur, D. *et al.* On the Immortality of Television Sets: 'Function' in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biol. Evol.* **5,** 578–590 (2013).

17. Niu, D.-K. & Jiang, L. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* **430,** 1340–1343 (2013).

18. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci.* **110,** 5294–5300 (2013).

19. Eddy, S. R. The ENCODE project: Missteps overshadowing a success. *Curr. Biol.* **23,** R259–R261 (2013).

20. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).

21. Fay, J. C., Wyckoff, G. J. & Wu, C.-I. Positive and Negative Selection on the Human Genome. *Genetics* **158,** 1227–1234 (2001).

22. Andolfatto, P. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437,** 1149–1152 (2005).

23. Eyre-Walker, A., Woolfit, M. & Phelps, T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics* **173,** 891–900 (2006).

24. Boyko, A. R. *et al.* Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLOS Genet.* **4,** e1000083 (2008).

25. Wilson, D. J., Hernandez, R. D., Andolfatto, P. & Przeworski, M. A Population Genetics-Phylogenetics Approach to Inferring Natural Selection in Coding Sequences. *PLOS Genet.* **7,** e1002395 (2011).

26. Ward, L. D. & Kellis, M. Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. *Science* **337,** 1675–1678 (2012).

27. Khurana, E. *et al.* Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science* **342,** 1235587 (2013).

28. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45,** 723–729 (2013).

29. Narlikar, L. *et al.* Genome-wide discovery of human heart enhancers. *Genome Res.* **20,** 381–392 (2010).

30. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11,** 294–296 (2014).

31. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28,** 817–825 (2010).

32. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).

33. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41,** 827–841 (2013).

34. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. *Mol. Biol. Evol.* **30,** 1159–1171 (2013).

35. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

36. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–1797 (2012).

37. Erwin, G. D. *et al.* Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLOS Comput. Biol.* **10,** e1003677 (2014).

38. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

39. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).

40. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).

41. Cooper, G. M. *et al.* Characterization of Evolutionary Rates and Constraints in Three Mammalian Genomes. *Genome Res.* **14,** 539–548 (2004).

42. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).

43. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

44. Ponting, C. P., Nellåker, C. & Meader, S. Rapid Turnover of Functional Sequence in Human and Other Genomes. *Annu. Rev. Genomics Hum. Genet.* **12,** 275–299 (2011).

45. Chiaromonte, F. *et al.* The Share of Human Genomic DNA under Selection Estimated from Human–Mouse Genomic Alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 245–254 (2003).

46. Meader, S., Ponting, C. P. & Lunter, G. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* **20,** 1335–1343 (2010).

47. Smith, N. G. C., Brandström, M. & Ellegren, H. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84,** 806–813 (2004).

48. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).

49. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLOS Genet.* **10,** e1004525 (2014).

50. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2016).

51. Lunter, G., Ponting, C. P. & Hein, J. Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model. *PLOS Comput. Biol.* **2,** e5 (2006).

52. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci.* **111,** 6131–6138 (2014).

53. Pheasant, M. & Mattick, J. S. Raising the estimate of functional human sequences. *Genome Res.* **17,** 1245–1253 (2007).

54. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43,** 1031–1034 (2011).

55. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

56. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. (Wiley-Interscience, 1991).

57. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinform.* **12,** 41–51 (2011).

58. Kondrashov, A. S. & Crow, J. F. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2,** 229–234 (1993).

59. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

60. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biol.* **16,** 56 (2015).

61. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Res.* **44,** D710–D716 (2016).

62. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).

63. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48,** 214–220 (2016).

64. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15,** 480 (2014).

65. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49,** 618–624 (2017).

66. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47,** 276–283 (2015).

67. Zhai, W., Nielsen, R. & Slatkin, M. An Investigation of the Statistical Power of Neutrality Tests Based on Comparative and Population Genetic Data. *Mol. Biol. Evol.* **26,** 273–283 (2009).

68. Project, the 1000 G. Variation in genome-wide mutation rates within and between human families. *Nat. Genet.* **43,** 712–714 (2011).

69. Rokach, L. & Maimon, O. Top-down induction of decision trees classifiers - a survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **35,** 476–487 (2005).

70. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136,** 665–677 (2017).

71. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16,** 22 (2015).

72. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

73. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–318 (2007).

74. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).

75. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol. Cell* **49,** 825–837 (2013).

76. Konopka, G. *et al.* Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature* **462,** 213–217 (2009).

77. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16,** 16–23 (2000).

78. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418,** 869–872 (2002).

79. Jorde, L. B. & Wooding, S. P. Genetic variation, classification and 'race'. *Nat. Genet.* **36,** S28–S33 (2004).

80. Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327,** 78–81 (2010).

81. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17,** 93–108 (2016).

82. Singleton, A. B., Hardy, J., Traynor, B. J. & Houlden, H. Towards a complete resolution of the genetic architecture of disease. *Trends Genet.* **26,** 438–442 (2010).

83. Előry, L., Halligan, D. L. & Keightley, P. D. Distributions of Selectively Constrained Sites and Deleterious Mutation Rates in the Hominid and Murid Genomes. *Mol. Biol. Evol.* **27,** 177–192 (2010).

84. Gulko, B. FitCons research homepage. *FitCons research homepage* (2017). Available at: www.fitcons.science.

85. UCSC Genomics Institute. Conservation Track Settings. *UCSC Genome Browser: Conservation Track* Available at: http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=cons100way. (Accessed: 1st August 2017)

86. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Comput. Biol.* **6,** e1001025 (2010).

87. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

88. Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE (Cap Analysis of Gene Expression): A Protocol for the Detection of Promoter and Transcriptional Networks. in *Gene Regulatory Networks* 181–200 (Humana Press, 2012). doi:10.1007/978-1-61779-292-2_11

89. Prescott, S. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163,** 68–83 (2015).

90. Ronan, J. L., Wu, W. & Crabtree, G. R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* **14,** 347–359 (2013).

91. Staahl, B. T. *et al.* Kinetic Analysis of npBAF to nBAF Switching Reveals Exchange of SS18 with CREST and Integration with Neural Developmental Pathways. *J. Neurosci.* **33,** 10348–10361 (2013).

92. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

93. Vattathil, S. & Akey, J. M. Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell* **163,** 281–284 (2015).

94. Vernot, B. & Akey, J. M. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* **343,** 1017–1021 (2014).

95. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31,** 761–763 (2015).

96. Timme, N., Alford, W., Flecker, B. & Beggs, J. M. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J. Comput. Neurosci.* **36,** 119–140 (2014).

97. Song, Q. *et al.* A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. *PLOS ONE* **8,** e81148 (2013).

98. Liu, F. *et al.* The Human Genomic Melting Map. *PLOS Comput. Biol.* **3,** e93 (2007).

99. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33,** 364–376 (2015).

100. Ernst, J. Roadmap 12 Mark, 25 Imputed Chromatin States - Emission Parameters. *Wustl Roadmap Project Web Site* (2013). Available at: http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/Imputed12Marks_25_States.pdf. (Accessed: 20th July 2017)

101. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34,** D590–D598 (2006).

102. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44,** D110–D115 (2016).